DELOCALIZATION FOR A CLASS OF RANDOM BLOCK BAND MATRICES

ZHIGANG BAO AND LÁSZLÓ ERDŐS

Abstract. We consider $N \times N$ Hermitian random matrices H consisting of blocks of size $M \geq N^{6/7}$. The matrix elements are i.i.d. within the blocks, close to a Gaussian in the four moment matching sense, but their distribution varies from block to block to form a block-band structure, with an essential band width M. We show that the entries of the Green's function $G(z) = (H - z)^{-1}$ satisfy the local semicircle law with spectral parameter $z = E + \mathrm{i}\eta$ down to the real axis for any $\eta \gg N^{-1}$, using a combination of the supersymmetry method inspired by [19] and the Green's function comparison strategy. Previous estimates were valid only for $\eta \gg M^{-1}$. The new estimate also implies that the eigenvectors in the middle of the spectrum are fully delocalized.

1. Introduction

The Hamiltonian of quantum systems on a graph Γ is a self-adjoint matrix $H = (h_{ab})_{a,b\in\Gamma}$, $H = H^*$. The matrix elements h_{ab} represent the quantum transition rates from vertex a to b. Disordered quantum systems have random matrix elements. We assume they are centered, $\mathbb{E}h_{ab} = 0$, and independent subject to the basic symmetry constraint $h_{ab} = \bar{h}_{ba}$. The variance $\sigma_{ab}^2 := \mathbb{E}|h_{ab}|^2$ represents the strength of the transition from a to b and we use a scaling where the norm $\|H\|$ is typically order 1. The simplest case is the mean field model, where h_{ab} are identically distributed; this is the standard Wigner matrix ensemble [29]. The other prominent example is the Anderson model [2] or random Schrödinger operator, $H = \Delta + V$, where the kinetic energy Δ is the (deterministic) graph Laplacian and the potential $V = (V_x)_{x\in\Gamma}$ is an on-site multiplication operator with random multipliers. If Γ is a discrete d-dimensional torus, then only few matrix elements h_{ab} are nonzero and they connect nearest neighbor points in the torus, $\mathrm{dist}(a, b) \leq 1$. This is in sharp contrast to the mean field character of the Wigner matrices.

Random band matrices naturally interpolate between the mean field Wigner matrices and the short range Anderson model. They are characterized by a parameter M, called the *band width*, such that the matrix elements h_{ab} for $\mathrm{dist}(a, b) \geq M$ are zero or negligible. If M is comparable with the diameter N of the system then we are in the mean field regime, while $M \sim 1$ corresponds to the short range model.

The Anderson model exhibits a metal-insulator phase transition: at high disorder the system is in the localized (insulator) regime, while at small disorder it is in the delocalized (metallic) regime, at least in $\mathsf{d} \geq 3$ dimensions and away from the spectral edges. The localized regime is characterized by exponentially decaying eigenfunctions and off diagonal decay of the Green's function, while in the complementary regime the eigenfunctions are supported in the whole physical space. In terms of the *localization length* ℓ, the characteristic length scale of the decay, the localized regime corresponds to $\ell \ll N$, while in the delocalized regime $\ell \sim N$. Starting from the basic papers [1, 15], the localized regime is well understood, but the delocalized regime is still an open mathematical problem for the d-dimensional torus.

Date: March 26, 2015.

2010 *Mathematics Subject Classification.* 15B52, 82B44.

Key words and phrases. Random band matrix, supersymmetry, Green's function comparison, local semicircle law, delocalization.

Z.G. Bao was supported by ERC Advanced Grant RANMAT No.338804; L. Erdős was partially supported by ERC Advanced Grant RANMAT No.338804 .

Since the eigenvectors of the mean field Wigner matrices are always delocalized [13, 14], while the short range models are localized, by varying the parameter M in the random band matrix, one expects a (de)localization phase transition. Indeed, for $\mathsf{d} = 1$ it is conjectured (and supported by non rigorous supersymmetric calculations [16]) that the system is delocalized for broad bands, $M \gg N^{1/2}$ and localized for $M \ll N^{1/2}$. The optimal power $1/2$ has not yet been achieved from either sides. Localization has been shown for $M \ll N^{1/8}$ in [22], while delocalization in a certain sense was proven for $M \gg N^{4/5}$ in [11]. Interestingly, for a special Gaussian model even the sine kernel behavior of the 2-point correlation function of the characteristic polynomials could be proven down to the optimal band width $M \gg N^{1/2}$, see [18, 20]. Note that the sine kernel is consistent with the delocalization but does not imply it. We remark that our discussion concerns the bulk of the spectrum; the transition at the spectral edge is much better understood. In [24] it was shown that the edge spectrum follows the Tracy Widom distribution, characteristic to mean field model, for $M \gg N^{5/6}$, but it yields a different distribution for narrow bands, $M \ll N^{5/6}$.

Delocalization is closely related to estimates on the diagonal elements of the resolvent $G(z) = (H-z)^{-1}$ at spectral parameters with small imaginary part $\eta = \mathsf{Im} z$. Indeed, if $G_{ii}(E + i\eta)$ is bounded for all i and all $E \in \mathbb{R}$, then each ℓ^2-normalized eigenvector $\mathbf{u}$ of H is delocalized on scale η^{-1} in a sense that $\max_i |u_i|^2 \lesssim \eta$, i.e. u is supported on at least η^{-1} sites. In particular, if G_{ii} can be controlled down to the scale $\eta \sim 1/N$, then the system is in the complete delocalized regime. Moreover, boundedness of G_{ii} also implies that the local semicircle law holds for the same regime of η.

For band matrices with band width M, or even under the more general condition $\sigma_{ab}^2 \le M^{-1}$, the boundedness of G_{ii} was shown down to scale $\eta \gg M^{-1}$ in [14] (see also [12]). If $M \gg N^{1/2}$, it is expected that G_{ii} remains bounded even down to $\eta \gg N^{-1}$ which is the typical eigenvalue spacing, the smallest relevant scale in the model. However, with the standard approach [14, 12] via the self-consistent equations for the Green's function does not seem to work for $\eta \le 1/M$; the fluctuation is hard to control. The more subtle approach using the self-consistent *matrix* equation in [11] could prove delocalization and the off-diagonal Green's function profile that are consistent with the conventional quantum diffusion picture, but it was valid only for relatively large η, far from M^{-1}. Moment methods, even with a delicate renormalization scheme [23] could not break the barrier $\eta \sim M^{-1}$ either.

In this paper we attack the problem differently; with supersymmetric (SUSY) techniques. Our main result is that $G_{ii}(z)$ is bounded, and the local semicircle law holds for any $\eta \gg N^{-1}$, i.e. down to the optimal scale, if the band width is not too small, $M \gg N^{6/7}$, but under two technical assumptions. First, we consider a generalization of Wegner's n-orbital model [21, 28], namely, we assume that the band matrix has a block structure, i.e. it consists of $M \times M$ blocks and the matrix elements within each block have the same distribution. This assumption is essential to reduce the number of integration variables in the supersymmetric representation, since, roughly speaking, each $M \times M$ block will be represented by a single supermatrix with 16 supersymmetric variables. Second, we assume that the distribution of the matrix elements matches a Gaussian up to four moments in the spirit of [27]. Supersymmetry heavily uses Gaussian integrations, in fact all mathematically rigorous works on random band matrices with supersymmetric method assume that the matrix elements are Gaussian, see [4, 5, 6, 18, 19, 20, 25, 26]. The Green's function comparison method [14] allows one to compare Green's functions of two matrix ensembles provided that the distributions match up to four moments and provided that G_{ii} are bounded. This was an important motivation to reach the optimal scale $\eta \gg N^{-1}$.

In the next subsections we introduce the model precisely and state our main results. While SUSY approach is ubiquitous in physics, see e.g. the basic monograph by Efetov [7], its application in rigorous proofs is notoriously difficult. Initiated by T. Spencer (see [25] for a summary) and starting with the

paper [4] by Disertori, Pinsker and Spencer, only a handful of mathematical papers have succeeded in exploiting this powerful tool. Our supersymmetric analysis was inspired by [19], but our observable, G_{ab}, requires a partly different formalism, in particular we use the singular version of the superbosonization formula [3]. Moreover, our analysis is considerably more involved since we consider relatively narrow bands. In Section 1.3, we explain our novelties compared with [19].

1.1. **Matrix model.** Let

$$H_N = (h_{ab})$$

be an $N \times N$ random Hermitian matrix, in which the entries are independent (up to symmetry), centered, complex variables. In this paper, we are concerned with H_N possessing a block band structure. To define this structure explicitly, we set the additional parameters $M \equiv M(N)$ and $W \equiv W(N)$ satisfying

$$W = N/M.$$

For simplicity, we assume that both M and W are integers. Let $S = (\mathfrak{s}_{jk})$ be a $W \times W$ symmetric matrix, which will be chosen as a weighted Laplacian of a connected graph on W vertices. Now, we decompose H_N into $W \times W$ blocks of size $M \times M$, and relabel

$$h_{jk,\alpha\beta} := h_{ab}, \qquad j,k = 1,\dots,W, \quad \alpha,\beta = 1,\dots,M,$$

where $(j,k) = (\lceil a/M \rceil, \lceil b/M \rceil)$ is the index of the block containing h_{ab}, and

$$(\alpha,\beta) = \Big(a-(j-1)M, b-(k-1)M\Big)$$

describes the location of the entry in the block. Moreover, we assume

$$\mathbb{E} h_{jk,\alpha\beta} h_{j'k',\alpha'\beta'} = \frac{1}{M}\delta_{jk'}\delta_{j'k}\delta_{\alpha\beta'}\delta_{\beta\alpha'}(\delta_{jk} + \mathfrak{s}_{jk}). \tag{1.1}$$

That means, the variance profile of the random matrix $\sqrt{M}H_N$ is given by

$$\tilde{S} = (\tilde{\mathfrak{s}}_{jk}) := I + S, \tag{1.2}$$

in which each entry represents the common variance of the entries in the corresponding block of $\sqrt{M}H_N$. Moreover, if $h_{jk,\alpha\beta}$'s are Gaussian, (1.1) also implies that for each off-diagonal entry $h_{jk,\alpha\beta}$, its real part and imaginary part are i.i.d. $N(0, \tilde{\mathfrak{s}}_{jk}/2M)$ variables.

1.2. **Assumptions and main results.** In the sequel, for some matrix $A = (a_{ij})$ and some index sets I and J, we introduce the notation $A^{(\mathsf{I}|\mathsf{J})}$ to denote the submatrix obtained by deleting the i-th row and j-th column of A for all $i \in \mathsf{I}$ and $j \in \mathsf{J}$. We will adopt the abbreviation

$$A^{(i|j)} := A^{(\{i\}|\{j\})}, \quad i \neq j, \qquad A^{(i)} := A^{(\{i\}|\{i\})}. \tag{1.3}$$

In addition, we use $||A||_{\max} := \max_{i,j}|a_{ij}|$ to denote the max norm of A. Throughout the paper, we need some assumptions on S.

Assumption 1.1 (On S)**.** *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a connected simple graph with $\mathcal{V} = \{1,\dots,W\}$. Assume that S is a $W \times W$ symmetric matrix satisfying the following four conditions.*

(i) *S is a weighted Laplacian on $\mathcal{G}$, i.e. for $i \neq j$, we have $\mathfrak{s}_{ij} > 0$ if $\{i,j\} \in \mathcal{E}$ and $\mathfrak{s}_{ij} = 0$ if $\{i,j\} \notin \mathcal{E}$, and for the diagonal entries, we have*

$$\mathfrak{s}_{ii} = -\sum_{j:j\neq i} \mathfrak{s}_{ij}, \qquad \forall\, i = 1,\dots,W.$$

(ii) *$\tilde{S}$ defined in (1.2) is strictly diagonally dominant, i.e., there exists some constant $c_0 > 0$ such that*

$$1 + 2\mathfrak{s}_{ii} > c_0, \qquad \forall\, i = 1, \dots, W.$$

(iii) *For the discrete Green's functions, we assume that there exist some positive constants C and γ such that*

$$\max_{i=1,\dots,W} ||(S^{(i)})^{-1}||_{\max} \le CW^{\gamma}.$$

(iv) *There exists a spanning tree $\mathcal{G}_0 = (\mathcal{V}, \mathcal{E}_0) \subset \mathcal{G}$, on which the weights are bounded below, i.e. for some constant $c > 0$, we have*

$$\mathfrak{s}_{ij} \ge c, \qquad \text{if} \quad \{i, j\} \in \mathcal{E}_0$$

Remark 1.2. *From Assumption 1.1 (ii), we easily see that*

$$\tilde{S} \ge c_0 I \tag{1.4}$$

for the same positive constant c_0. In addition, the lower bound c in (iv) can be weakened to $N^{-\varepsilon}$ for some sufficiently small constant $\varepsilon > 0$. But for simplicity, we will not try to optimize this bound in this paper.

Example 1.1. *Let Δ be the standard discrete Laplacian on the d-dimensional torus $[1, \mathfrak{w}]^{\mathsf{d}} \cap \mathbb{Z}^{\mathsf{d}}$, with periodic boundary condition, where $\mathfrak{w} = W/\mathsf{d}$. Here by standard we mean the weights on the edges of the box are all 1. Now let $S = a\Delta$ for some positive constant $a < 1/4\mathsf{d}$. It is then easy to check Assumption 1.1 (i), (ii) and (iv) are satisfied. In addition, if $\mathsf{d} = 1$, it is well known that we can choose $\gamma = 1$ in Assumption 1.1 (iii). For $\mathsf{d} \ge 3$, one can choose $\gamma = 0$. For $\mathsf{d} = 2$, one can choose $\gamma = \varepsilon$ for arbitrarily small constant ε. For instance, one can refer to [8] for more details.*

For simplicity, we also introduce the notation

$$\sigma_{ij}^2 := \mathbb{E}|h_{ij}|^2, \qquad \mathcal{T} := (\sigma_{ij}^2)_{N,N} = \frac{1}{M}\tilde{S} \otimes \mathbf{1}_M \mathbf{1}_M', \quad i, j = 1, \dots, N, \tag{1.5}$$

where $\mathbf{1}_M$ is the M-dimensional vector whose components are all 1 and $\tilde{S}$ is the variance matrix in (1.2). It is elementary that

$$\mathrm{Spec}(\mathcal{T}) = \mathrm{Spec}(\tilde{S}) \cup \{0\} \subset [0, 1]. \tag{1.6}$$

Our assumption on M depends on the constant γ in Assumption 1.1 (iii).

Assumption 1.3 (On M)**.** *We assume that there exists a (small) positive constant ε_1 such that*

$$M \ge W^{4+2\gamma+\varepsilon_1}. \tag{1.7}$$

Remark 1.4. *A direct consequence of (1.7) and $N = MW$ is*

$$M \ge N^{\frac{4+2\gamma+\varepsilon_1}{5+2\gamma+\varepsilon_1}}. \tag{1.8}$$

Especially, when $\gamma = 1$, one has $M \gg N^{6/7}$. Actually, through a more involved analysis, (1.7) (or (1.8)) can be further improved. At least, for $\gamma \le 1$, we expect that $M \gg N^{4/5}$ is enough. However, we will not pursue this direction here.

Besides Assumption 1.1 on the variance profile of H, we need to impose some additional assumption on the distribution of its entries. To this end, we temporarily employ the notation $H^g = (h_{ab}^g)$ to represent a random block band matrix with Gaussian entries, satisfying (1.1), Assumption 1.1 and Assumption 1.3.

Assumption 1.5 (On distribution). *We assume that for each $a, b \in \{1, \dots, N\}$, the moments of the entry h_{ab} match those of h^g_{ab} up to the 4th order, i.e.*

$$\mathbb{E}(\mathsf{Re}h_{ab})^k(\mathsf{Im}h_{ab})^\ell = \mathbb{E}(\mathsf{Re}h^g_{ab})^k(\mathsf{Im}h^g_{ab})^\ell, \quad \forall\, k, \ell \in \mathbb{N}, \quad s.t. \quad k + \ell \le 4. \tag{1.9}$$

In addition, we assume the distribution of h_{ab} possesses a subexponential tail, namely, there exist positive constants c_1 and c_2 such that for any $\tilde{\gamma} > 0$,

$$\mathbb{P}\Big(|h_{ab}| \ge \tilde{\gamma}^{c_1}(\mathbb{E}|h_{ab}|^2)^{\frac{1}{2}}\Big) \le c_2 e^{-\tilde{\gamma}} \tag{1.10}$$

holds uniformly for all $a, b = 1, \dots, N$.

The four moment condition (1.9) in the context of random matrices first appeared in Tao and Vu's work [27].

To state our results, we will need the following notion on the comparison of two random sequences, which was introduced in [9] and [12].

Definition 1.6 (Stochastic domination). *For some possibly N-dependent parameter set U_N, and two families of random variables $\mathsf{X} = (\mathsf{X}_N(u) : N \in \mathbb{N}, u \in \mathsf{U}_N)$ and $\mathsf{Y} = (\mathsf{Y}_N(u) : N \in \mathbb{N}, u \in \mathsf{U}_N)$, we say that X is stochastically dominated by Y, if for all $\varepsilon > 0$ and $D > 0$ we have*

$$\sup_{u \in \mathsf{U}_N} \mathbb{P}\Big(\mathsf{X}_N(u) \ge N^\varepsilon \mathsf{Y}_N(u)\Big) \le N^{-D} \tag{1.11}$$

for all sufficiently large $N \ge N_0(\varepsilon, D)$. In this case we write

$$\mathsf{X} \prec \mathsf{Y}.$$

For example, by (1.1) and Assumption 1.5, we have

$$|h_{ab}| \prec \frac{1}{\sqrt{M}}, \quad \forall\, a, b = 1, \dots, N. \tag{1.12}$$

Note that $\tilde{S}$ is doubly stochastic. It is known that the empirical eigenvalue distribution of H_N converges to the semicircle law, whose density function is given by

$$\varrho_{sc}(x) := \frac{1}{2\pi}\sqrt{4 - x^2} \cdot \mathbf{1}(|x| \le 2).$$

We denote the Green's function of H_N by

$$G(z) \equiv G_N(z) := (H_N - z)^{-1}, \quad z = E + \mathbf{i}\eta \in \mathbb{C}^+ := \{w \in \mathbb{C} : \mathsf{Im}w > 0\}$$

and its (a, b) matrix element is $G_{ab}(z)$. Throughout the paper, we will always use E and η to denote the real and imaginary part of z without further mention. In addition, for simplicity, we suppress the subscript N from the notation of the matrices here and there. The Stieltjes transform of $\varrho_{sc}(x)$ is

$$m_{sc}(z) = \int_{-2}^{2} \frac{\varrho_{sc}(x)}{x - z}\mathrm{d}x = \frac{-z + \sqrt{z^2 - 4}}{2},$$

where we chose the branch of the square root with positive imaginary part for $z \in \mathbb{C}^+$. Note that $m_{sc}(z)$ is a solution to the following self-consistent equation

$$m_{sc}(z) = \frac{1}{-z - m_{sc}(z)}. \tag{1.13}$$

The semicircle law also holds in a local sense, see Theorem 2.3 in [12]. For simplicity, we cite this result with a slight modification adjusted to our assumption.

Proposition 1.7 (Erdős, Knowles, Yau, Yin, [12]). *Let H be a random block band matrix satisfying Assumptions 1.1, 1.3 and 1.5. Then*

$$\max_{a,b}|G_{ab}(z)-\delta_{ab}m_{sc}(z)|\prec\frac{1}{\sqrt{M\eta}},\quad \text{if}\quad E\in[-2+\kappa,2-\kappa]\quad\text{and}\quad M^{-1+\varepsilon}\le\eta\le 10 \tag{1.14}$$

for any fixed small positive constants κ and ε.

Remark 1.8. *We remark that Theorem 2.3 in [12] was established under a more general assumption $\sum_k\sigma_{jk}^2=1$ and $\sigma_{jk}^2\le C/M$. Especially, the block structure on the variance profile is not needed. In addition, Theorem 2.3 in [12] also covers the edges of the spectrum, which will not be discussed in this paper. We also refer to [14] for a previous result, see Theorem 2.1 therein.*

Our aim in this paper is to extend the local semicircle law to the regime $\eta\gg N^{-1}$ and replace M with N in (1.14). More specifically, we will work in the following set, defined for arbitrarily small constant $\kappa>0$ and any sufficiently small positive constant $\varepsilon_2:=\varepsilon_2(\varepsilon_1)$,

$$\mathbf{D}(N,\kappa,\varepsilon_2):=\Big\{z=E+\mathbf{i}\eta\in\mathbb{C}:|E|\le\sqrt{2}-\kappa,N^{-1+\varepsilon_2}\le\eta\le M^{-1}N^{\varepsilon_2}\Big\}. \tag{1.15}$$

Throughout the paper, we will assume that ε_2 is much smaller than ε_1, see (1.7) for the latter. Specifically, there exists some large enough constant C such that $\varepsilon_2\le\varepsilon_1/C$.

Theorem 1.9 (Local semicircle law). *Suppose that H is a random block band matrix satisfying Assumptions 1.1, 1.3 and 1.5. Let κ be an arbitrarily small positive constant and ε_2 be any sufficiently small positive constant. Then*

$$\max_{a,b}|G_{ab}(z)-\delta_{ab}m_{sc}(z)|\prec\frac{1}{\sqrt{N\eta}} \tag{1.16}$$

holds uniformly on $\mathbf{D}(N,\kappa,\varepsilon_2)$.

Remark 1.10. *The restriction $|E|\le\sqrt{2}-\kappa$ in (1.15) is technical. We believe the result can be extended to the whole bulk regime of the spectrum, i.e., $|E|\le 2-\kappa$, see Section 12 for further comment. The upper bound of η in (1.15) is also technical. However, for $\eta>M^{-1}N^{\varepsilon_2}$, one can control the Green's function by (1.14) directly.*

Let $\lambda_1,\dots,\lambda_N$ be the eigenvalues of H_N. We denote by $\mathbf{u}_i:=(u_{i1},\dots,u_{iN})$ the normalized eigenvector of H_N corresponding to λ_i. From Theorem 1.9, we can also get the following delocalization property for the eigenvectors.

Theorem 1.11 (Complete delocalization). *Let H be a random block band matrix satisfying Assumptions 1.1, 1.3 and 1.5. We have*

$$\max_{i:|\lambda_i|\le\sqrt{2}-\kappa}||\mathbf{u}_i||_\infty\prec N^{-\frac{1}{2}}. \tag{1.17}$$

Remark 1.12. *We remark that delocalization in a certain weak sense was proven in [11] for an even more general class of random band matrices if $M\gg N^{4/5}$. However, Theorem 1.11 asserts delocalization for* **all** *eigenvectors in a very strong sense (supremum norm), while Proposition 7.1 of [11] stated that* **most** *eigenvectors are delocalized in a sense that their substantial support cannot be too small.*

1.3. **Outline of the proof strategy and novelties.** In this section, we briefly outline the strategy for the proof of Theorem 1.9.

The first step, which is the main task of the whole proof, is to establish the following Theorem 1.14, namely, a prior estimate of the Green's function in the Gaussian case. For technical reason, we need the following slight modification of Assumption 1.3, to state the result.

Assumption 1.13 (On M). *Let ε_1 be the small positive constant in Assumption 1.3. We assume*

$$N(\log N)^{-10} \geq M \geq W^{4+2\gamma+\varepsilon_1}. \tag{1.18}$$

In the regime $M \geq N(\log N)^{-10}$, we see that (1.16) anyway follows from (1.14) directly.

Theorem 1.14. *Assume that H is a Gaussian block band matrix, satisfying Assumptions 1.1 and 1.13. Let n be any fixed positive integer. Let κ be an arbitrarily small positive constant and ε_2 be any sufficiently small positive constant. There is $N_0 = N_0(n)$, such that for all $N \geq N_0$ and all $z \in \mathbf{D}(N, \kappa, \varepsilon_2)$, we have*

$$\mathbb{E}|G_{ab}(z)|^{2n} \leq N^{C_0}\Big(\delta_{ab} + \frac{1}{(N\eta)^n}\Big), \quad \forall\, a, b = 1, \ldots, N \tag{1.19}$$

for some positive constant C_0 independent of n and z.

Remark 1.15. *Much more delicate analysis can show that the prefactor N^{C_0} can be improved to some n-dependent constant C_n. We refer to Section 12 for further comment on this issue.*

Using the definition of stochastic domination in Definition 1.6, a simple Markov inequality shows that (1.19) implies

$$|G_{ab}(z)| \prec \delta_{ab} + \frac{1}{\sqrt{N\eta}}, \quad \forall\, a, b = 1, \ldots, N. \tag{1.20}$$

The proof of Theorem 1.14 is the main task of our paper. We will use the supersymmetry method. We partially rely on the arguments from Shcherbina's work [19] concerning universality of the local 2-point function and we develop new techniques to treat our observable, the high moment of the entries of $G(z)$, under a more general setting. We will comment on the novelties later in this subsection.

The second step is to generalize Theorem 1.14 from the Gaussian case to more general distribution satisfying Assumption 1.5, via a Green's function comparison strategy initiated in [14], see Lemma 2.1 below.

The last step is to use Lemma 2.1 and its Corollary 2.2 to prove our main theorems. Using (1.20) below to bound the error term in the self-consistent equation for the Green's function, we can prove Theorem 1.9 by a continuity argument in z, with the aid of the initial estimate for large η provided in Proposition 1.7. Theorem 1.11 will then easily follow from Theorem 1.9.

The second and the last steps are carried out in Section 2. The main body of this paper, Sections 3–11 is devoted to the proof of Theorem 1.14.

One of the main novelty of this work is to combine the supersymmetry method and the Green's function comparison strategy to go beyond the Gaussian ensemble, which was so far the only random band matrix ensemble amenable to the supersymmetry method, as mentioned at the beginning. The comparison strategy requires an apriori control on the individual matrix elements of the Green's function with high probability (see (1.20)), this is one of our main motivations behind Theorem 1.14.

Although we consider a different observable than [19], many technical aspects of the supersymmetric analysis overlaps with [19]. For the convenience of the reader, we now briefly introduce the strategy of [19], and highlight the main novelties of our work.

In [19], the author considers the 2-point correlation function of the trace of the resolvent of the Gaussian block band matrix H, with the variance profile $\tilde{S} = 1 + a\Delta$, under the assumption $M \sim N$ (note that we use M instead of W in [19] for the size of the blocks). The 2-point correlation function can be expressed in terms of a superintegral of a superfunction $F(\{\breve{\mathcal{S}}_i\}_{i=1}^W)$ with a collection of 4×4 supermatrices $\breve{\mathcal{S}}_i := \mathcal{Z}_i^* \mathcal{Z}_i$. Here for each i, $\mathcal{Z}_i = (\Psi_{1,i}, \Psi_{2,i}, \Phi_{1,i}, \Phi_{2,i})$ is an $M \times 4$ matrix and $\mathcal{Z}_i^*$ is its conjugate transpose, where $\Psi_{1,i}$ and $\Psi_{2,i}$ are Grassmann M-vectors whilst $\Phi_{1,i}$ and $\Phi_{2,i}$ are complex M-vectors. Then, by using the superbosonization formula in the nonsingular case ($M \geq 4$) from [17],

one can transform the superintegral of $F(\{\breve{\mathcal{S}}_i\}_{i=1}^W)$ to a superintegral of $F(\{\mathcal{S}_i\}_{i=1}^W)$, where each $\mathcal{S}_i$ is a supermatrix akin to $\breve{\mathcal{S}}_i$, but only consists of 16 independent variables (either complex or Grassmann). We will call the integral representation of the observable after using the superbosonization formula as the *final integral representation*. Schematically it has the form

$$\int \mathsf{g}(\mathcal{S}_c) e^{M\mathsf{f}_c(\mathcal{S}_c)+\mathsf{f}_g(\mathcal{S}_g,\mathcal{S}_c)} \mathrm{d}\mathcal{S}, \tag{1.21}$$

for some functions $\mathsf{g}(\cdot)$, $\mathsf{f}_c(\cdot)$ and $\mathsf{f}_g(\cdot)$, where we used the abbreviation $\mathcal{S} := \{\mathcal{S}_i\}_{i=1}^W$ and $\mathcal{S}_c$ and $\mathcal{S}_g$ represents the collection of all complex variables and Grassmann variables in $\mathcal{S}$, respectively. Here, $\mathsf{g}(\mathcal{S}_c)$ and $\mathsf{f}_c(\mathcal{S}_c)$ are some complex functions and $\mathsf{f}_g(\mathcal{S}_g, \mathcal{S}_c)$ will be mostly regarded as a function of the Grassmann variables with complex variables as its parameters. The number of variables (either complex or Grassmann) in the final integral representation then turns out to be of order W, which is much smaller than the original order N. In fact, in [19] it is assumed that $W = O(1)$ although the author also mentions the possibility to deal with the case $W \sim N^\varepsilon$ for some small positive ε, see the remark below Theorem 1 therein.

Performing a saddle point analysis for the complex measure $\exp\{M\mathsf{f}_c(\mathcal{S}_c)\}$, one can restrict the integral in a small vicinity of some saddle point, say, $\mathcal{S}_c = \mathcal{S}_{c0}$. It turns out that $\mathsf{f}_c(\mathcal{S}_{c0}) = 0$ and $\mathsf{f}_c(\mathcal{S}_c)$ decays quadratically away from $\mathcal{S}_{c0}$. Consequently, by plugging in the saddle point $\mathcal{S}_{c0}$, one can estimate $\mathsf{g}(\mathcal{S}_c)$ by $\mathsf{g}(\mathcal{S}_{c0})$ directly. However, for $\exp\{M\mathsf{f}_c(\mathcal{S}_c)\}$ and $\exp\{\mathsf{f}_g(\mathcal{S}_g, \mathcal{S}_c)\}$, one shall expand them around the saddle point. Roughly speaking, in some vicinity of $\mathcal{S}_{c0}$, one will find that the expansions read

$$e^{M\mathsf{f}_c(\mathcal{S}_c)} = \exp\{-\mathbf{u}'\mathbb{A}\mathbf{u} + \mathsf{e}_c(\mathbf{u})\}, \qquad e^{\mathsf{f}_g(\mathcal{S}_g,\mathcal{S}_c)} = \exp\{-\boldsymbol{\rho}'\mathbb{H}\boldsymbol{\tau}\}\mathsf{p}(\boldsymbol{\rho}, \boldsymbol{\tau}, \mathbf{u}), \tag{1.22}$$

where $\mathbf{u}$ is a complex vector of dimension $O(W)$, which is essentially a vectorization of $\sqrt{M}(\mathcal{S}_c - \mathcal{S}_{c0})$; $\mathsf{e}_c(\mathbf{u}) = o(1)$ is some error term; $\boldsymbol{\rho}$ and $\boldsymbol{\tau}$ are two Grassmann vectors of dimension $O(W)$; $\mathbb{A}$ is a complex matrix with positive-definite Hermitian part and $\mathbb{H}$ is a complex matrix; $\mathsf{p}(\boldsymbol{\rho}, \boldsymbol{\tau}, \mathbf{u})$ is the expansion of $\exp\{\mathsf{f}_g(\mathcal{S}_g, \mathcal{S}_c) - \mathsf{f}_g(\mathcal{S}_g, \mathcal{S}_{c0})\}$, which possesses the form

$$\mathsf{p}(\boldsymbol{\rho}, \boldsymbol{\tau}, \mathbf{u}) = \sum_{\ell=0}^{O(W)} M^{-\frac{\ell}{2}} \mathsf{p}_\ell(\boldsymbol{\rho}, \boldsymbol{\tau}, \mathbf{u}), \tag{1.23}$$

where $\mathsf{p}_\ell(\boldsymbol{\rho}, \boldsymbol{\tau}, \mathbf{u})$ is a polynomial of the components of $\boldsymbol{\rho}$ and $\boldsymbol{\tau}$ with degree 2ℓ, regarding $\mathbf{u}$ as fixed parameters. Now, keeping the leading order term of $\mathsf{p}(\boldsymbol{\rho}, \boldsymbol{\tau}, \mathbf{u})$, and discarding the remainder terms, one can get the final estimate of the integral by taking the Gaussian integral over $\mathbf{u}$, $\boldsymbol{\rho}$ and $\boldsymbol{\tau}$. This completes the summary of [19].

Similarly to [19], we also use the superbosonization formula to reduce the number of variables and perform the saddle point analysis on the resulting integral. However, owing to the following three main aspects, our analysis is significantly different from [19].

•(Different observable) Our objective is to compute high moments of the single entry of the Green's function. By using Wick's formula (see Proposition 3.1), we express $\mathbb{E}|G_{jk}|^{2n}$ in terms of a superintegral of some superfunction of the form

$$\tilde{F}\Big(\{\Psi_{a,j}, \Psi^*_{a,j}, \Phi_{a,j}, \Phi^*_{a,j}\}_{\substack{a=1,2;\\ j=1,\dots,W}}\Big) := \big(\bar{\phi}_{1,q,\beta}\phi_{1,p,\alpha}\bar{\phi}_{2,p,\alpha}\phi_{2,q,\beta}\big)^n F(\{\breve{\mathcal{S}}_i\}_{i=1}^W)$$

for some $p, q \in \{1, \dots, W\}$ and $\alpha, \beta \in \{1, \dots, M\}$, where $\phi_{1,p,\alpha}$ is the α-th coordinate of $\Phi_{1,p}$, and the others are defined analogously. Unlike the case in [19], $\tilde{F}$ is not a function of $\{\breve{\mathcal{S}}_i\}_{i=1}^W$ only. Hence, using the superbosonization formula to change $\breve{\mathcal{S}}_i$ to $\mathcal{S}_i$ directly is not feasible in our case. In order to handle the factor $\big(\bar{\phi}_{1,q,\beta}\phi_{1,p,\alpha}\bar{\phi}_{2,p,\alpha}\phi_{2,q,\beta}\big)^n$, the main idea is to split off certain rank-one supermatrices from $\breve{\mathcal{S}}_p$ and $\breve{\mathcal{S}}_q$ such that this factor can be expressed in terms of the entries of these rank-one supermatrices.

Then we use the superbosonization formula not only in the nonsingular case from [17] but also in the singular case from [3] to change and reduce the variables, resulting the final integral representation of $\mathbb{E}|G_{jk}|^{2n}$. Though this final integral representation, very schematically, is still of the form (1.21), due to the decomposition of the supermatrices $\breve{\mathcal{S}}_p$ and $\breve{\mathcal{S}}_q$, it is considerably more complicated than its counterpart in [19]. Especially, the function $\mathsf{g}(\mathcal{S}_c)$ differs from its counterpart in [19], and its estimate at the saddle point follows from a different argument.

•(Small band width) In [19], the author considers the case that the band width M is comparable with N, i.e. the number of blocks W is finite. Though the derivation of the 2-point correlation function is highly nontrivial even with such a large band width, our objective, the local semicircle law and delocalization of the eigenvectors, however, can be proved for the case $M \sim N$ in a similar manner as for the Wigner matrix ($M = N$), see [12, 14]. In our work, we will work with much smaller band width to go beyond the results in [12, 14], see Assumption 1.3. Several main difficulties stemming from a narrow band width can be heuristically explained as follows.

At first, let us focus on the integral over the small vicinity of the saddle point, in which the exponential functions in the integrand in (1.21) approximately look like (1.22).

We regard the first term in (1.22) as a complex Gaussian measure, of dimension $O(W)$. When $W \sim 1$, one can discard the error term $\mathsf{e}_c(\mathbf{u})$ directly and perform the Gaussian integral over $\mathbf{u}$, due to the fact $\int \mathrm{d}\mathbf{u} \exp\{-\mathbf{u}'\mathsf{Re}(\mathbb{A})\mathbf{u}\}|\mathsf{e}_c(\mathbf{u})| = o(1)$. However, such an estimate is not allowed when $W \sim N^{\varepsilon}$ (say), because the normalization of the measure $\exp\{-\mathbf{u}'\mathsf{Re}(\mathbb{A})\mathbf{u}\}$ might be exponentially larger than that of $\exp\{-\mathbf{u}'\mathbb{A}\mathbf{u}\}$. In order to handle this issue, we shall do a second deformation of the contours of the complex variables in the vicinity of the saddle, following the steepest descent paths exactly, whereby we can transform the complex Gaussian measure to a real one, thus the error term of the integral can be controlled.

Now, we turn to the second term in (1.22). When $W \sim 1$, there are only finitely many Grassmann variables. Hence, the complex coefficient of each term in the polynomial $\mathsf{p}(\boldsymbol{\rho}, \boldsymbol{\tau}, \mathbf{u})$, which is of order $M^{-\ell/2}$ for some $\ell \in \mathbb{N}$ (see (1.23)), actually controls the magnitude of the integral of this term against the Gaussian measure $\exp\{-\boldsymbol{\rho}'\mathbb{H}\boldsymbol{\tau}\}$. Consequently, in case of $W \sim 1$, it suffices to keep the leading order term (according to $M^{-\ell/2}$), one may discard the others trivially, and compute the Gaussian integral over $\boldsymbol{\rho}$ and $\boldsymbol{\tau}$ explicitly. However, when $W \sim N^{\varepsilon}$ (say), in light of the Wick's formula (3.2) and the fact that the coefficients are of order $M^{-\ell/2}$, the order of the integral of each term of $\mathsf{p}(\boldsymbol{\rho}, \boldsymbol{\tau}, \mathbf{u})$ against the Gaussian measure reads $M^{-\ell/2} \det \mathbb{H}^{(\mathsf{I}|\mathsf{J})}$ for some index sets I and J and some $\ell \in \mathbb{N}$. Due to the fact $W \sim N^{\varepsilon}$, $\det \mathbb{H}^{(\mathsf{I}|\mathsf{J})}$ is typically exponential in W. Hence, it is much more complicated to determine and compare the orders of the integrals of all $e^{O(W)}$ terms. In our discussion, we perform a unified estimate for the integrals of all the terms, rather than simply compare them by $M^{-\ell/2}$.

In addition, the analysis for the integral away from the vicinity of the saddle point in our work is also quite different from [19]. Actually, the integral over the complement of the vicinity can be trivially ignored in [19], since each factor in the integrand of (1.21) is of order 1, thus gaining any $o(1)$ factor for the integrand outside the vicinity is enough for the estimate. However, in our case, either $\exp\{M\mathsf{f}_c(\mathcal{S}_c)\}$ or $\int \mathrm{d}\mathcal{S}_g \exp\{\mathsf{f}_g(\mathcal{S}_g, \mathcal{S}_c)\}$ is essentially exponential in W. This fact forces us to provide an apriori bound for $\int \mathrm{d}\mathcal{S}_g \exp\{\mathsf{f}_g(\mathcal{S}_g, \mathcal{S}_c)\}$ in the full domain of $\mathcal{S}_c$ rather than in the vicinity of the saddle point only. In addition, an analysis of the tail behavior of the measure $\exp\{M\mathsf{f}_c(\mathcal{S}_c)\}$ needs also to be performed.

•(General variance profile $\tilde{S}$) In [19], the authors considered the special case $S = a\Delta$ with $a < 1/4\mathsf{d}$. We generalize the discussion to more general weighted Laplacians S satisfying Assumption 1.1, which, as a special case, includes the standard Laplacian Δ for any fixed dimension d.

1.4. **Notation and organization.** Throughout the paper, we will need some notation. At first, we conventionally use $U(r)$ to denote the unitary group of degree r, as well, $U(1,1)$ represents the $U(1,1)$ group. Furthermore, we denote

$$\mathring{U}(r) = U(r)/U(1)^r, \qquad \mathring{U}(1,1) = U(1,1)/U(1)^2.$$

Recalling the real part E of z, we will frequently need the following two parameters

$$a_+ = \frac{\mathbf{i}E + \sqrt{4-E^2}}{2}, \qquad a_- = \frac{\mathbf{i}E - \sqrt{4-E^2}}{2}.$$

Correspondingly, we define the following four matrices

$$D_\pm = \mathrm{diag}(a_+, a_-), \quad D_\mp = \mathrm{diag}(a_-, a_+), \quad D_+ = \mathrm{diag}(a_+, a_+), \quad D_- = \mathrm{diag}(a_-, a_-). \tag{1.24}$$

We remark here $D_\pm$ does not mean "D_+ or D_-". For simplicity, we introduce the following notation for some domains used throughout the paper.

$$\mathbb{I} := [0,1], \quad \mathbb{L} := [0, 2\pi), \quad \Sigma : \quad \text{unit circle}, \quad \mathbb{R}_+ := [0,\infty), \quad \mathbb{R}_- := -\mathbb{R}_+, \quad \Gamma := a_+\mathbb{R}_+.$$

For some $\ell \times \ell$ Hermitian matrix A, we use $\lambda_1(A) \le \ldots \le \lambda_\ell(A)$ to represent its ordered eigenvalues. For some possibly N-dependent parameter set U_N, and two families of complex functions $\{a_N(u) : N \in \mathbb{N}, u \in \mathsf{U}_N\}$ and $\{b_N(u) : N \in \mathbb{N}, u \in \mathsf{U}_N\}$, if there exists a positive constant $C > 1$ such that $C^{-1}|b_N(u)| \le |a_N(u)| \le C|b_N(u)|$ holds uniformly in N and u, we write

$$a_N(u) \sim b_N(u).$$

Conventionally, we use $\{\mathbf{e}_i : i = 1, \ldots, \ell\}$ to denote the standard basis of $\mathbb{R}^\ell$, in which the dimension ℓ has been suppressed for simplicity. For some real quantities a and b, we use $a \wedge b$ and $a \vee b$ to represent $\min\{a, b\}$ and $\max\{a, b\}$, respectively.

Throughout the paper, c, c', c_1, c_2, C, C', C_1, C_2 represent some generic positive constants that are possibly n-dependent and may differ from line to line. In contrast, we use C_0 to denote some generic positive constant independent of n.

The paper will be organized in the following way. In Section 2, we prove Theorem 1.9 and Theorem 1.11, with Theorem 1.14. The proof of Theorem 1.14 will be done in Section 3–Section 11. More specifically, in Section 3, we use the supersymmetric formalism to represent $\mathbb{E}|G_{ij}|^{2n}$ in terms of a superintegral, in which the integrand can be factorized into several functions; Section 4 is devoted to a preliminary analysis on these functions; Section 5–Section 10 are responsible for different steps of the saddle point analysis, whose organization will be further clarified at the end of Section 5; Section 11 is devoted to the final proof of Theorem 1.14, by summing up the discussions in 3–Section 10. Finally, in Section 12, we make some further comments on possible improvements.

2. Proofs of Theorem 1.9 and Theorem 1.11

At first, (1.19) can be generalized to the generally distributed matrix with the four moment matching condition via the Green's function comparison strategy.

Lemma 2.1. *Assume that H is a random block band matrix, satisfying Assumptions 1.1, 1.5 and 1.13. Let κ be an arbitrarily small positive constant and ε_2 be any sufficiently small positive constant. There is $N_0 = N_0(n)$, such that for all $N \geq N_0$ and all $z \in \mathbf{D}(N, \kappa, \varepsilon_2)$, we have*

$$\mathbb{E}|G_{ab}(z)|^{2n} \leq N^{C_0}\Big(\delta_{ab} + \frac{1}{(N\eta)^n}\Big), \quad \forall\, a, b = 1, \ldots, N \tag{2.1}$$

for some positive constant C_0 uniform in n and z.

By the definition of stochastic domination in Definition 1.6, we can get the following corollary from Lemma 2.1 immediately.

Corollary 2.2. *Under the assumptions of Lemma 2.1, we have*

$$|G_{ab}(z)| \prec \delta_{ab} + \frac{1}{\sqrt{N\eta}}, \quad \forall\, a, b = 1, \ldots, N \tag{2.2}$$

uniformly on $\mathbf{D}(N, \kappa, \varepsilon_2)$.

In the sequel, at first, we prove Lemma 2.1 from Theorem 1.14 via the Green's function comparison strategy. Then we prove Theorem 1.9, using Lemma 2.1. Finally, we will show that Theorem 1.11 follows from Theorem 1.9 simply.

2.1. **Green's function comparison: Proof of Lemma 2.1.** To show (2.1), we use Lindeberg's replacement strategy to compare the Green's functions of the Gaussian case and the general case. That means, we will replace the entries of H^g by those of H one by one, and compare the Green's functions step by step. Choose and fix a bijective ordering map

$$\varpi : \{(i, j) : 1 \leq i \leq j \leq N\} \to \Big\{1, \ldots, \varsigma(N)\Big\}, \quad \varsigma(N) := \frac{N(N+1)}{2}. \tag{2.3}$$

Then we use H_k to represent the $N \times N$ random Hermitian matrix whose (i, j)-th entry is h_{ij} if $\varpi(i, j) \leq k$, and is h^g_{ij} otherwise. Especially, we have $H_0 = H^g$ and $H_{\varsigma(N)} = H$. Correspondingly, we define the Green's functions by

$$G_k(z) := \Big(H_k - z\Big)^{-1}, \quad k = 1, \ldots, \varsigma(N).$$

Fix k and denote

$$\varpi^{-1}(k) = (a, b). \tag{2.4}$$

Then, we write

$$H_{k-1} = H_k^0 + \mathsf{V}_{ab}, \qquad \mathsf{V}_{ab} := \Big(1 - \frac{\delta_{ab}}{2}\Big)\big(h^g_{ab}\mathbf{e}_a\mathbf{e}_b^* + h^g_{ba}\mathbf{e}_b\mathbf{e}_a^*\big),$$
$$H_k = H_k^0 + \mathsf{W}_{ab}, \qquad \mathsf{W}_{ab} := \Big(1 - \frac{\delta_{ab}}{2}\Big)\big(h_{ab}\mathbf{e}_a\mathbf{e}_b^* + h_{ba}\mathbf{e}_b\mathbf{e}_a^*\big),$$

where H_k^0 is obtained via replacing h_{ab} and h_{ba} by 0 in H_k (or replacing h^g_{ab} and h^g_{ba} by 0 in H_{k-1}). In addition, we denote

$$G_k^0(z) = (H_k^0 - z)^{-1}.$$

Set $\varepsilon_3 \equiv \varepsilon_3(\gamma, \varepsilon_1)$ to be a sufficiently small positive constant, satisfying (say)

$$\varepsilon_3 \leq \frac{1}{100} \cdot \frac{\varepsilon_1}{5 + 2\gamma + \varepsilon_1}, \tag{2.5}$$

where γ is from Assumption 1.1 (iii) and ε_1 is from (1.7). For simplicity, we introduce the following parameters for $\ell = 1, \dots, \varsigma(N)$ and $i, j = 1, \dots, N$,

$$\widehat{\Theta}_0 := N^{C_0}, \quad \widehat{\Theta}_{\ell,ij} := \widehat{\Theta}_0 \Big(1 + C\Big(\frac{N^{\varepsilon_3}}{\sqrt{M}}\Big)^5\Big)^{\ell} \prod_{\varpi(a,b)\le \ell} \Big(1 + C\delta_{\{i,j\}\{a,b\}}\Big(\frac{N^{\varepsilon_3}\sqrt{N\eta}}{\sqrt{M}}\Big)^5\Big), \tag{2.6}$$

where C is a positive constant. Here we used the notation $\delta_{\mathsf{IJ}} = 1$ if two index sets I and J are the same and $\delta_{\mathsf{IJ}} = 0$ otherwise. It is easy to see that for $\eta \le M^{-1}N^{\varepsilon_2}$, we have

$$\widehat{\Theta}_{\ell,ij} \le 2\widehat{\Theta}_0, \qquad \forall\, \ell = 1, \dots, \varsigma(N), \quad i, j = 1, \dots, N, \tag{2.7}$$

by using (1.8).

Now, we compare $G_{k-1}(z)$ and $G_k(z)$. We will prove the following lemma.

Lemma 2.3. *Suppose that the assumptions in Lemma 2.1 hold. Additionally, we assume that for some sufficiently small positive constant ε_3 satisfying (2.5),*

$$|(G_\ell)_{ij}(z)| \prec N^{\varepsilon_3}, \quad |(G^0_\ell)_{ij}(z)| \prec N^{\varepsilon_3}, \quad \forall\, \ell = 1, \dots, \varsigma(N), \quad \forall\, i, j = 1, \dots, N \tag{2.8}$$

uniformly for $z \in \mathbf{D}(N, \kappa, \varepsilon_2)$. Let $n \in \mathbb{N}$ be any given integer. Then, if

$$\mathbb{E}|(G_{k-1})_{ij}(z)|^{2n} \le \widehat{\Theta}_{k-1,ij}\Big(\delta_{ij} + \frac{1}{(N\eta)^n}\Big), \quad \forall\, i, j = 1, \dots, N, \tag{2.9}$$

we also have

$$\mathbb{E}|(G_k)_{ij}(z)|^{2n} \le \widehat{\Theta}_{k,ij}\Big(\delta_{ij} + \frac{1}{(N\eta)^n}\Big), \qquad \forall\, i, j = 1, \dots, N \tag{2.10}$$

for any $k = 1, \dots, \varsigma(N)$.

Proof of Lemma 2.3. Fix k and omit the argument z from now on. At first, under the conditions (2.8) and (2.9), we show that

$$\mathbb{E}|(G^0_k)_{ij}|^{2n} \le 3\widehat{\Theta}_0\Big(\delta_{ij} + \frac{1}{(N\eta)^n}\Big), \quad \forall\, i, j = 1, \dots, N. \tag{2.11}$$

To see this, we use the expansion with (2.4)

$$(G^0_k)_{ij} = (G_{k-1})_{ij} + (G_{k-1}\mathsf{W}_{ab}G^0_k)_{ij},$$

which implies that for a sufficiently large constant $D > 0$

$$\mathbb{E}|(G^0_k)_{ij}|^{2n} \le \mathbb{E}\Big||(G_{k-1})_{ij}| + \frac{N^{2\varepsilon_3}}{\sqrt{M}}\Big|^{2n} + \eta^{-2n}N^{-D} \le 3\widehat{\Theta}_0\Big(\delta_{ij} + \frac{1}{(N\eta)^n}\Big),$$

where the first step follows from (1.12), (2.8), Definition 1.6 and the trivial bound η^{-1} for the Green's functions, and the second step follows from (2.9), (2.7) and the fact $N^{2\varepsilon_3}/\sqrt{M} \ll 1/\sqrt{N\eta}$ for $z \in \mathbf{D}(N, \kappa, \varepsilon_2)$.

Now, recall (2.4) again and expand $G_{k-1}(z)$ and $G_k(z)$ around $G^0_k(z)$, namely

$$\begin{aligned} G_{k-1} &= G^0_k + \sum_{\ell=1}^{m}(-1)^\ell(G^0_k\mathsf{V}_{ab})^\ell G^0_k + (-1)^{m+1}(G^0_k\mathsf{V}_{ab})^{m+1}G_{k-1}, \\ G_k &= G^0_k + \sum_{\ell=1}^{m}(-1)^\ell(G^0_k\mathsf{W}_{ab})^\ell G^0_k + (-1)^{m+1}(G^0_k\mathsf{W}_{ab})^{m+1}G_k. \end{aligned} \tag{2.12}$$

We always choose m to be sufficiently large, depending on ε_3 but independent of N. Then, we can write

$$(G_{k-1})_{ij} = (G^0_k)_{ij} + \sum_{\ell=1}^{m}\mathsf{R}_{\ell,ij} + \tilde{\mathsf{R}}_{m+1,ij},$$

$$(G_k)_{ij} = (G_k^0)_{ij} + \sum_{\ell=1}^{m} \mathsf{S}_{\ell,ij} + \tilde{\mathsf{S}}_{m+1,ij}, \tag{2.13}$$

where

$$\begin{aligned} \mathsf{R}_{\ell,ij} &:= (-1)^\ell \Big((G_k^0 \mathsf{V}_{ab})^\ell G_k^0 \Big)_{ij}, \quad \mathsf{S}_{\ell,ij} := (-1)^\ell \Big((G_k^0 \mathsf{W}_{ab})^\ell G_k^0 \Big)_{ij}, \quad \ell = 1, \dots, m, \\ \tilde{\mathsf{R}}_{m+1,ij} &:= (-1)^{m+1} \Big((G_k^0 \mathsf{V}_{ab})^{m+1} G_{k-1} \Big)_{ij}, \quad \tilde{\mathsf{S}}_{m+1,ij} := (-1)^{m+1} \Big((G_k^0 \mathsf{W}_{ab})^{m+1} G_k \Big)_{ij}. \end{aligned} \tag{2.14}$$

At first, by taking m sufficiently large, from (2.8) and (1.12), we have the trivial bound

$$|\tilde{\mathsf{R}}_{m+1,ij}|, |\tilde{\mathsf{S}}_{m+1,ij}| \prec M^{-\frac{m+1}{2}} N^{(m+2)\varepsilon_3} \ll \frac{1}{M^3 \sqrt{N\eta}}. \tag{2.15}$$

For $\mathsf{R}_{\ell,ij}$ and $\mathsf{S}_{\ell,ij}$, we split the discussion into off-diagonal case and diagonal case. In the case of $i \neq j$, we keep the first and the last factors of the terms in the expansions of $((G_k^0 \mathsf{V}_{ab})^\ell G_k^0)_{ij}$ and $((G_k^0 \mathsf{W}_{ab})^\ell G_k^0)_{ij}$, namely, $(G_k^0)_{ij'}$ and $(G_k^0)_{i'j}$ for some $i', j' = a$ or b, and bound the factors in between by using (1.12) and (2.8), resulting the bound

$$|\mathsf{R}_{\ell,ij}|, |\mathsf{S}_{\ell,ij}| \prec M^{-\frac{\ell}{2}} N^{(\ell-1)\varepsilon_3} \sum_{i',j'=a,b} |(G_k^0)_{ij'} (G_k^0)_{i'j}|, \quad \ell = 1, \dots, m. \tag{2.16}$$

For $i = j$, we only keep the first factor of the terms in the expansions of $((G_k^0 \mathsf{V}_{ab})^\ell G_k^0)_{ii}$ and $((G_k^0 \mathsf{W}_{ab})^\ell G_k^0)_{ii}$, and bound the others by using (1.12) and (2.8), resulting the bound

$$|\mathsf{R}_{\ell,ii}|, |\mathsf{S}_{\ell,ii}| \prec M^{-\frac{\ell}{2}} N^{\ell\varepsilon_3} \Big(|(G_k^0)_{ia}| + |(G_k^0)_{ib}| \Big), \quad \ell = 1, \dots, m. \tag{2.17}$$

Observe that, in case $i \neq j$, if $\{i,j\} \neq \{a,b\}$, at least one of $(G_k^0)_{ij'}$ and $(G_k^0)_{i'j}$ is an off-diagonal entry of G_k^0 for $i', j' = a$ or b.

Now we compare the $2n$-th moment of $|(G_{k-1})_{ij}|$ and $|(G_k)_{ij}|$. At first, we write

$$\mathbb{E}|(G_d)_{ij}|^{2n} = \mathbb{E}((G_d)_{ij})^n (\overline{(G_d)_{ij}})^n, \quad d = k-1, k \tag{2.18}$$

By substituting the expansion (2.13) into (2.18), we can write

$$\mathbb{E}|(G_d)_{ij}|^{2n} = \mathbf{A}(i,j) + \mathbf{R}_d(i,j), \quad d = k-1, k, \tag{2.19}$$

where $\mathbf{A}(i,j)$ is the sum of the terms which depend only on H_k^0 and the first four moments of h_{ab}, and $\mathbf{R}_d(i,j)$ is the sum of all the other terms. We claim that $\mathbf{R}_d(i,j)$ satisfies the bound

$$|\mathbf{R}_d(i,j)| \leq C \widehat{\Theta}_0 \Big(\frac{N^{\varepsilon_3}}{\sqrt{M}} \Big)^5 \Big(\delta_{ij} + \frac{\delta_{\{i,j\}\{a,b\}}}{(N\eta)^{n-\frac{5}{2}}} + \frac{1}{(N\eta)^n} \Big), \qquad d = k-1, k, \tag{2.20}$$

for some positive constant C. Now, we verify (2.20). According to (2.11) and the fact that the sequence $\mathsf{R}_{1,ij}, \dots, \mathsf{R}_{m,ij}, \tilde{\mathsf{R}}_{m+1,ij}$, as well as $\mathsf{S}_{1,ij}, \dots, \mathsf{S}_{m,ij}, \tilde{\mathsf{S}}_{m+1,ij}$, decreases by a factor $N^{\varepsilon_3}/\sqrt{M}$ in magnitude, it is not difficult to check the leading order terms of $\mathbf{R}_{k-1}(i,j)$ are of the form

$$\mathbb{E} \left((G_k^0)_{ij} \right)^p \left(\overline{(G_k^0)_{ij}} \right)^{2n-p-\sum_{\ell=1}^5 (q_\ell + q'_\ell)} \prod_{\ell=1}^5 \mathsf{R}_{\ell,ij}^{q_\ell} \bar{\mathsf{R}}_{\ell,ij}^{q'_\ell}, \tag{2.21}$$

and those of $\mathbf{R}_k(i,j)$ are of the form

$$\mathbb{E} \left((G_k^0)_{ij} \right)^p \left(\overline{(G_k^0)_{ij}} \right)^{2n-p-\sum_{\ell=1}^5 (q_\ell + q'_\ell)} \prod_{\ell=1}^5 \mathsf{S}_{\ell,ij}^{q_\ell} \bar{\mathsf{S}}_{\ell,ij}^{q'_\ell}, \tag{2.22}$$

with some $p, q_\ell, q'_\ell \in \mathbb{N}$ such that

$$\sum_{\ell=1}^5 \ell(q_\ell + q'_\ell) = 5, \qquad 0 \leq p \leq 2n - \sum_{\ell=1}^5 (q_\ell + q'_\ell). \tag{2.23}$$

Every other term has at least 6 factors of h_{ab} or h^g_{ab} or their conjugates, thus their sizes are typically controlled by $M^{-3}(N\eta)^{-n}$, i.e. they are subleading. Hence, it suffices to bound (2.21) and (2.22). In the sequel, we only estimate (2.21) in details, (2.22) can be handled in the same manner.

Now, the five factors of h_{ab} or h_{ba} within the $\mathsf{R}_{\ell,ij}$'s in (2.21) are independent of the rest and estimated by $M^{-5/2}$. For the remaining factors from G^0_k, we use (2.11) to bound $2n$ of them and use (2.8) to bound the rest. In the case that $i\neq j$ and $\{i,j\}\neq\{a,b\}$, by the discussion above, we must have an off-diagonal entry of G^0_k in the product $(G^0_k)_{ij'}(G^0_k)_{i'j}$ for any choice of $i',j'=a$ or b. Then, in the bound for $\mathsf{R}_{\ell,ij}$ in (2.16), for each $(G^0_k)_{ij'}(G^0_k)_{i'j}$, we keep the off-diagonal entry and bound the other by N^{ε_3} from assumption (2.8). Hence, by using (2.16) and (2.23), we see that for some $i_r,j_r\in\{i,j,a,b\}$ with $i_r\neq j_r$, $r=1,\dots,\sum(q_\ell+q'_\ell)$, the following bound holds

$$(2.21)\le\Big(\frac{N^{\varepsilon_3}}{\sqrt{M}}\Big)^5\mathbb{E}\Big(|(G^0_k)_{ij}|^{2n-\sum_{\ell=1}^5(q_\ell+q'_\ell)}\prod_{r=1}^{\sum_{\ell=1}^5(q_\ell+q'_\ell)}|(G^0_k)_{i_rj_r}|\Big)\le 3\Big(\frac{N^{\varepsilon_3}}{\sqrt{M}}\Big)^5\frac{\widehat{\Theta}_0}{(N\eta)^n},\tag{2.24}$$

where the last step follows from (2.11) and Hölder's inequality. In case of $i\neq j$ but $\{i,j\}=\{a,b\}$, we keep an entry in the product $(G^0_k)_{ij'}(G^0_k)_{i'j}$ and bound the other by N^{ε_3}. We remark here in this case the entry being kept can be either diagonal or off-diagonal. Consequently, for some $i_r,j_r\in\{i,j,a,b\}, r=1,\dots,\sum(q_\ell+q'_\ell)$, we have the bound

$$(2.21)\le\Big(\frac{N^{\varepsilon_3}}{\sqrt{M}}\Big)^5\mathbb{E}\Big(|(G^0_k)_{ij}|^{2n-\sum_{\ell=1}^5(q_\ell+q'_\ell)}\prod_{r=1}^{\sum_{\ell=1}^5(q_\ell+q'_\ell)}|(G^0_k)_{i_rj_r}|\Big)\le 3\Big(\frac{N^{\varepsilon_3}}{\sqrt{M}}\Big)^5\frac{\widehat{\Theta}_0}{(N\eta)^{n-\frac52}}\tag{2.25}$$

by using (2.11) and Hölder's inequality again. Hence, we have shown (2.20) in the case of $i\neq j$. For $i=j$, it is analogous to show

$$(2.21)\le 3\Big(\frac{N^{\varepsilon_3}}{\sqrt{M}}\Big)^5\widehat{\Theta}_0\tag{2.26}$$

by using (2.11), (2.17) and Hölder's inequality. Hence, we verified (2.20). Consequently, by Assumption 1.5, (2.19) and (2.20) we have

$$\Big|\mathbb{E}|(G_{k-1})_{ij}|^{2n}-\mathbb{E}|(G_k)_{ij}|^{2n}\Big|\le C\widehat{\Theta}_0\Big(\frac{N^{\varepsilon_3}}{\sqrt{M}}\Big)^5\Big(\delta_{ij}+\frac{\delta_{\{i,j\}\{a,b\}}}{(N\eta)^{n-\frac52}}+\frac{1}{(N\eta)^n}\Big),$$

which together with the assumption (2.9) for $\mathbb{E}|(G_{k-1})_{ij}|^{2n}$ and the definition of $\widehat{\Theta}_{\ell,ij}$'s in (2.6), we can get

$$\mathbb{E}|(G_k)_{ij}|^{2n}\le\widehat{\Theta}_{k,ij}\Big(\delta_{ij}+\frac{1}{(N\eta)^n}\Big).$$

Hence, we completed the proof of Lemma 2.3. □

To show (2.1), we also need the following lemma.

Lemma 2.4. *Suppose that the assumptions in Lemma 2.1 hold. Fix the indices $a,b\in\{1,\dots N\}$. Let H^0 be a matrix obtained from H with its (a,b)-th entry replaced by 0. Then, if for some $\eta_0\ge 1/N$ there exists*

$$|G_{ii}(z)|\prec 1,\quad |(G^0)_{ii}(z)|\prec 1\quad \text{for}\quad \eta\ge\eta_0,\quad \forall\, i=1,\dots,N,\tag{2.27}$$

then we also have

$$|G_{ij}(z)|\prec\frac{\eta_0}{\eta},\quad |(G^0)_{ij}(z)|\prec\frac{\eta_0}{\eta},\quad \text{for}\quad \frac{1}{N}<\eta\le\eta_0,\quad \forall\, i,j=1,\dots,N.$$

Proof of Lemma 2.4. The proof is almost the same as the discussion on pages 2311–2312 in [10]. For the convenience of the reader, we sketch it below. At first, according to the discussion below (4.28) in [10], for any $i, j = 1, \dots, N$, we have

$$|G_{ij}(E + \mathbf{i}\eta)| \le C \max_{\ell} \sum_{k \ge 0} \mathsf{Im} G_{\ell\ell}(E + \mathbf{i}2^k \eta).$$

Now, we set

$$k_1 := \max\{k : 2^k \eta < \eta_0\}, \qquad k_2 := \max\{k : 2^k \eta < 1\}.$$

According to our assumption, both k_1 and k_2 are of the order $\log N$. Now, we have

$$\begin{aligned} \sum_{k \ge 0} \mathsf{Im} G_{\ell\ell}(E + \mathbf{i}2^k \eta) &= \sum_{k=0}^{k_1} \mathsf{Im} G_{\ell\ell}(E + \mathbf{i}2^k \eta) + \sum_{k=k_1}^{k_2} \mathsf{Im} G_{\ell\ell}(E + \mathbf{i}2^k \eta) + \sum_{k=k_2+1}^{\infty} \mathsf{Im} G_{\ell\ell}(E + \mathbf{i}2^k \eta) \\ &\prec \frac{\eta_0}{\eta} \sum_{k=0}^{k_1} \frac{1}{2^k} \mathsf{Im} G_{\ell\ell}(E + \mathbf{i}\eta_0) + (k_2 - k_1) + 1 \prec \frac{\eta_0}{\eta} \end{aligned}$$

where in the second step, we used the fact that the function $y \mapsto y \mathsf{Im} G_{\ell\ell}(E + \mathbf{i}y)$ is monotonically increasing, the condition (2.27) and the fact $\eta \le \eta_0$. Hence, we conclude the proof of Lemma 2.4. □

Now, with Theorem 1.14, Lemma 2.3 and Lemma 2.4, we can prove Lemma 2.1.

Proof for Lemma 2.1. The proof relies on the following bootstrap argument, namely, we show that once

$$|(G_\ell)_{ij}| \prec 1, \quad \forall\, \ell = 1, \dots, \varsigma(N), \quad \forall\, i, j = 1, \dots, N \tag{2.28}$$

holds for $\eta \ge \eta_0$ with $\eta_0 \in [N^{-1+\varepsilon_2+\varepsilon_3}, M^{-1}N^{\varepsilon_2}]$, it also holds for $\eta \ge \eta_0 N^{-\varepsilon_3}$ for any ε_3 satisfying (2.5). Assuming (2.28) holds for $\eta \ge \eta_0$, we see that

$$\begin{aligned} \max_{i,j} |(G_\ell^0)_{ij}| &= \max_{i,j} |(G_\ell)_{ij} + ((G_\ell)\mathsf{W}_{ab} G_\ell^0)_{ij}| \\ &\prec \max_{i,j} |(G_\ell)_{ij}| + \frac{1}{\sqrt{M}} \max_{i,j} |(G_\ell)_{ij}| \cdot \max_{i,j} |(G_\ell^0)_{ij}| \prec 1 + \frac{1}{\sqrt{M}} \max_{i,j} |(G_\ell^0)_{ij}|. \end{aligned}$$

Consequently, for $\eta \ge \eta_0$, we also have

$$|(G_\ell^0)_{ij}| \prec 1, \quad \forall\, \ell = 1, \dots, \varsigma(N), \quad \forall\, i, j = 1, \dots, N. \tag{2.29}$$

Therefore, (2.27) holds. Then, by Lemma 2.4, we see that (2.8) holds for $\eta \ge \eta_0 N^{-\varepsilon_3}$. Furthermore, by Lemma 2.3 and Theorem 1.14 for G_0, i.e. the Gaussian case, one can get that for any given n,

$$\begin{aligned} \mathbb{E}|(G_\ell)_{ij}|^{2n} \le \widehat{\Theta}_{\ell,ij}\Big(\delta_{ij} + \tfrac{1}{(N\eta)^n}\Big) \le 2\widehat{\Theta}_0\Big(\delta_{ij} + \tfrac{1}{(N\eta)^n}\Big), \quad \text{for} \quad M^{-1}N^{\varepsilon_2} \ge \eta \ge \eta_0 N^{-\varepsilon_3}, \\ \forall\, \ell = 1, \dots, \varsigma(N), \quad \forall\, i, j = 1, \dots, N. \end{aligned} \tag{2.30}$$

Note that since (2.30) holds for any given n, we get (2.28) for $M^{-1}N^{\varepsilon_2} \ge \eta \ge \eta_0 N^{-\varepsilon_3}$.

Now we start from $\eta_0 = M^{-1}N^{\varepsilon_2}$. By Proposition 1.7 we see that (2.28) holds for all $\eta \ge \eta_0$. Then we can use the bootstrap argument above finitely many times to show (2.28) holds for all $\eta \ge N^{-1+\varepsilon_2}$. Consequently, we have (2.8) for all $\eta \ge N^{-1+\varepsilon_2}$. Then, Lemma 2.1 follows from Lemma 2.3 and Theorem 1.14 immediately. □

2.2. **Proof of Theorem 1.9.** Without loss of generality we can assume that $M \leq N(\log N)^{-10}$, otherwise, Proposition 1.7 implies (1.16) immediately. Now, recalling the notation defined in (1.3), we denote the Green's function of $H^{(i)}$ as

$$G^{(i)}(z) := (H^{(i)} - z)^{-1},$$

with a little abuse of notation. We only need to consider the diagonal entries G_{ii} below, since the bound for the off-diagonal entires of $G(z)$ is implied by (2.1) directly. Set

$$\Delta_i \equiv \Delta_i(z) := \frac{1}{G_{ii}} + z + \sum_a \sigma_{ai}^2 G_{aa}. \tag{2.31}$$

We introduce the notation

$$\Lambda_d \equiv \Lambda_d(z) := \max_i |G_{ii}(z) - m_{sc}(z)|.$$

We have the following lemma.

Lemma 2.5. *Suppose that H satisfies Assumptions 1.1, 1.5 and 1.13. We have*

$$|\Delta_i(z)| \prec \frac{1}{\sqrt{N\eta}}, \quad \forall\, i = 1, \ldots, N, \tag{2.32}$$

uniformly for $z \in \mathbf{D}(N, \kappa, \varepsilon_2)$.

The proof of Lemma 2.5 will be postponed. Using Lemma 2.5, we see that, with high probability, (2.31) is a small perturbation of the self-consistent equation of m_{sc}, i.e. (1.13), considering $\sum_a \sigma_{ai}^2 = 1$. To control Λ_d, we use a *continuity argument* from [12].

We remind here that in the sequel, the parameter set of the stochastic dominance is always $\mathbf{D}(N, \kappa, \varepsilon_2)$, without further mention. We need to show that

$$\Lambda_d \prec \frac{1}{\sqrt{N\eta}}, \tag{2.33}$$

and first we claim that it suffices to show that

$$\mathbf{1}(\Lambda_d \leq N^{-\frac{\varepsilon_2}{4}})\Lambda_d \prec \frac{1}{\sqrt{N\eta}}. \tag{2.34}$$

Indeed, if (2.34) were proven, we see that with high probability either $\Lambda_d > N^{-\frac{\varepsilon_2}{4}}$ or $\Lambda_d \prec 1/\sqrt{N\eta} \leq N^{-\frac{\varepsilon_2}{2}}$ for $z \in \mathbf{D}(N, \kappa, \varepsilon_2)$. That means, there is a gap in the possible range of Λ_d. Now, choosing ε in (1.14) to be sufficiently small, we are able to get for $\eta = M^{-1}N^{\varepsilon_2}$,

$$\Lambda_d \prec N^{-\frac{\varepsilon_2}{2}}, \quad \forall\, E \in [-2 + \kappa, 2 - \kappa], \quad \forall\, i = 1, \ldots, N. \tag{2.35}$$

By the fact that Λ_d is continuous in z, we see that with high probability, Λ_d can only stay in one side of the range, namely, (2.33) holds. The rigorous details of this argument involve considering a fine discrete grid of the z-parameter and using that $G(z)$ is Lipschitz continuous (albeit with a large Lipschitz constant $1/\eta$). The details are found in Section 5.3 of [12].

Hence, what remains is to verify (2.34). The proof of (2.34) is almost the same as that for Lemma 3.5 in [14]. For the convenience of the reader, we sketch it below without reproducing the details. We set

$$\bar{m} \equiv \bar{m}(z) := \frac{1}{N}\sum_{i=1}^N G_{ii}(z), \quad \mathsf{u}_i \equiv \mathsf{u}_i(z) := G_{ii} - \bar{m}, \quad i = 1, \ldots, N.$$

We also denote $\vec{\mathsf{u}} := (\mathsf{u}_1, \ldots, \mathsf{u}_N)$. By the assumption $\Lambda_d \leq N^{-\frac{\varepsilon_2}{4}}$, we have

$$\mathsf{u}_i = O(N^{-\frac{\varepsilon_2}{4}}). \tag{2.36}$$

Now we rewrite (2.31) as

$$0 = G_{ii} + \frac{1}{z + \sum_a \sigma_{ai}^2 G_{aa} - \Delta_i} =: G_{ii} + \frac{1}{z + \bar{m}(z)} + \Omega_i. \tag{2.37}$$

By using (2.32), Lemma 5.1 in [14], and the assumption $\Lambda_d \leq N^{-\frac{\varepsilon_2}{4}}$, we can show that

$$\Omega_i = -\frac{\sum_a \sigma_{ai}^2 \mathsf{u}_a}{(z + \bar{m}(z))^2} + O\big(||\vec{\mathsf{u}}||_\infty^2\big) + O\big(\max_i |\Delta_i|\big). \tag{2.38}$$

One can refer to the derivation of (5.14) in [14] for more details. Averaging over i for (2.37) and (2.38) leads to

$$\bar{m}(z) + \frac{1}{z + \bar{m}(z)} = -\Omega. \tag{2.39}$$

and

$$\Omega := \frac{1}{N}\sum_{i=1}^N \Omega_i = O\big(||\vec{\mathsf{u}}||_\infty^2\big) + O\big(\max_i |\Delta_i|\big) \tag{2.40}$$

Plugging (2.36) and (2.32) into (2.40) yields

$$|\Omega| \prec N^{-\frac{\varepsilon_2}{2}}. \tag{2.41}$$

Using (2.41), the fact $|\bar{m}(z) - m_{sc}(z)| \leq \Lambda_d \leq N^{-\frac{\varepsilon_2}{4}}$, and Lemma 5.2 in [14], to (2.39), we have

$$|\bar{m}(z) - m_{sc}(z)| \leq |\Omega| = O\big(||\vec{\mathsf{u}}||_\infty^2\big) + O\big(\max_i |\Delta_i|\big), \tag{2.42}$$

where in the first step we have used the fact that $z \in \mathbf{D}(N, \kappa, \varepsilon_2)$ thus away from the edges of the semicircle law. Now, we combine (2.37), (2.38) and (2.39), resulting

$$\mathsf{u}_i = \frac{\sum_a \sigma_{ai}^2 \mathsf{u}_a}{(z + \bar{m}(z))^2} + \Omega + O\big(||\vec{\mathsf{u}}||_\infty^2\big) + O\big(\max_i |\Delta_i|\big) = \mathsf{w}_i + \frac{\sum_a \sigma_{ai}^2 \mathsf{u}_a}{(z + m_{sc}(z))^2}, \qquad i = 1, \ldots, N. \tag{2.43}$$

We just take the above identity as the definition of w_i. Analogously, we set $\vec{\mathsf{w}} := (\mathsf{w}_1, \ldots, \mathsf{w}_N)'$. Then (2.40) and (2.43) imply

$$\begin{aligned} ||\vec{\mathsf{w}}||_\infty &= O\big(||\vec{\mathsf{u}}||_\infty^2\big) + O\big(\max_i |\Delta_i|\big) + O\big(||\vec{\mathsf{u}}||_\infty \cdot |(z + \bar{m}(z))^{-2} - (z + m_{sc}(z))^{-2}|\big) \\ &\leq O\big(||\vec{\mathsf{u}}||_\infty^2\big) + O\big(\max_i |\Delta_i|\big) + O\big(||\vec{\mathsf{u}}||_\infty \cdot |\bar{m}(z) - m_{sc}(z)|\big) \\ &\leq O\big(||\vec{\mathsf{u}}||_\infty^2\big) + O\big(\max_i |\Delta_i|\big) \end{aligned} \tag{2.44}$$

where the second step follows from the fact $|z + m_{sc}(z)| \geq 1$ in $\mathbf{D}(N, \kappa, \varepsilon_2)$ (see (5.1) in [14] for instance), (2.41) and (2.42), and in the last step we used (2.42) again.

Now, using the fact $m_{sc}^2(z) = (m_{sc}(z) + z)^{-2}$ (see (1.13)), we rewrite (2.43) in terms of the matrix $\mathcal{T}$ introduced in (1.5) as

$$\vec{\mathsf{u}} = \big(1 - m_{sc}^2(z)\mathcal{T}\big)^{-1} \vec{\mathsf{w}}.$$

Consequently, we have

$$||\vec{\mathsf{u}}||_\infty \leq \Big|\Big|\big(1 - m_{sc}^2(z)\mathcal{T}\big)^{-1}\Big|\Big|_{\ell^\infty \to \ell^\infty} ||\vec{\mathsf{w}}||_\infty := \Gamma(z)||\vec{\mathsf{w}}||_\infty. \tag{2.45}$$

Then for $z \in \mathbf{D}(N, \kappa, \varepsilon_2)$, using (1.6) and Proposition A.2 (ii) in [12] (with $\delta_- = 1$ and $\theta > c$), we can get

$$\Gamma(z) = O(\log N). \tag{2.46}$$

Plugging (2.46) and (2.44) into (2.45) yields

$$||\vec{\mathsf{u}}||_\infty \prec O(||\vec{\mathsf{u}}||_\infty^2 + O\big(\max_i |\Delta_i|\big)) \prec ||\vec{\mathsf{u}}||_\infty^2 + \frac{1}{\sqrt{N\eta}},$$

where the second step follows from (2.32). Then (2.36) further implies that

$$||\vec{\mathbf{u}}||_\infty \prec \frac{1}{\sqrt{N\eta}},$$

which together with (2.42) and (2.32) also implies

$$|\bar{m}(z) - m_{sc}(z)| \prec \frac{1}{\sqrt{N\eta}}.$$

Hence

$$\Lambda_d \le ||\vec{\mathbf{u}}||_\infty + |\bar{m}(z) - m_{sc}(z)| \prec \frac{1}{\sqrt{N\eta}}.$$

Therefore, we completed the proof of Theorem 1.9.

Proof of Lemma 2.5. For simplicity, we omit the variable z from the notation below. At first, we recall the elementary identity by Schur's complement, namely,

$$G_{ii} = \frac{1}{h_{ii} - z - (\mathbf{h}_i^{\langle i\rangle})^* G^{(i)} \mathbf{h}_i^{\langle i\rangle}}. \tag{2.47}$$

where we used the notation $\mathbf{h}_i^{\langle i\rangle}$ to denote the i-th column of H, with the i-th component deleted. Now, we use the identity for $a, b \neq i$ (see Lemma 4.5 in [12] for instance),

$$G_{ab}^{(i)} = G_{ab} - G_{ai}G_{ib}(G_{ii})^{-1} = G_{ab} - G_{ai}G_{ib}\Big(h_{ii} - z - (\mathbf{h}_i^{\langle i\rangle})^* G^{(i)} \mathbf{h}_i^{\langle i\rangle}\Big). \tag{2.48}$$

By using (1.10) and the large deviation estimate for the quadratic form (see Theorem C.1 of [12] for instance), we have

$$\Big|(\mathbf{h}_i^{\langle i\rangle})^* G^{(i)} \mathbf{h}_i^{\langle i\rangle} - \sum_{a\neq i} \sigma_{ai}^2 \cdot G_{aa}^{(i)}\Big| \prec \sqrt{\frac{1}{M}\max_a |G_{aa}^{(i)}|^2 + \max_{a\neq b} |G_{ab}^{(i)}|^2}, \tag{2.49}$$

which implies that

$$\Big|(\mathbf{h}_i^{\langle i\rangle})^* G^{(i)} \mathbf{h}_i^{\langle i\rangle}\Big| \prec \max_{a\neq i} |G_{aa}^{(i)}| + \sqrt{\frac{1}{M}\max_{a\neq i} |G_{aa}^{(i)}|^2 + \max_{a\neq b} |G_{ab}^{(i)}|^2} \le 3\max_{a,b\neq i} |G_{ab}^{(i)}|, \tag{2.50}$$

where we have used the fact that $\sum_a \sigma_{ai}^2 = 1$ in the first inequality above. Plugging (1.20) and (2.50) into (2.48) and using Corollary 2.2 we obtain

$$\max_{a,b\neq i} |G_{ab}^{(i)}| \prec 1 + \frac{1}{N\eta}\Big(1 + 3\max_{a,b\neq i} |G_{ab}^{(i)}|\Big),$$

which implies

$$\max_{a,b\neq i} |G_{ab}^{(i)}| \prec 1, \qquad \Big|(\mathbf{h}_i^{\langle i\rangle})^* G^{(i)} \mathbf{h}_i^{\langle i\rangle}\Big| \prec 1. \tag{2.51}$$

In addition, (1.20), (2.48) and (2.51) lead to the fact that

$$|G_{ab}(z) - G_{ab}^{(i)}(z)| \prec \frac{1}{N\eta}, \qquad |G_{ab}^{(i)}(z)| \prec \delta_{ab} + \frac{1}{\sqrt{N\eta}}, \quad \forall\, a, b \neq i. \tag{2.52}$$

Now, using (2.31), (2.47), (2.49) and (2.52), we can see that

$$|\Delta_i| = \Big| - h_{ii} + (\mathbf{h}_i^{\langle i\rangle})^* G^{(i)} \mathbf{h}_i^{\langle i\rangle} - \sum_a \sigma_{ai}^2 G_{aa}\Big| \prec \frac{1}{\sqrt{N\eta}}. \tag{2.53}$$

Therefore, we completed the proof of Lemma 2.5. □

2.3. **Proof of Theorem 1.11.** With Theorem 1.9, we can prove Theorem 1.11 routinely. At first, due to Definition 1.6 and the fact that $G_{ab}(z)$ and $m_{sc}(z)$ are Lipschitz functions of z with Lipschitz constant η^{-1}, it is easy to strengthen (1.16) to

$$\max_{a,b} \sup_{z\in\mathbf{D}(N,\kappa,\varepsilon_2)} |G_{ab}(z) - \delta_{ab} m_{sc}(z)| \prec \frac{1}{\sqrt{N\eta}},$$

which implies that

$$\max_{a} \sup_{z\in\mathbf{D}(N,\kappa,\varepsilon_2)} |G_{aa}(z)| \prec C \tag{2.54}$$

for some positive constant C due to the fact that $m_{sc}(z) \sim 1$. Recalling the normalized eigenvector $\mathbf{u}_i = (u_{i1}, \ldots, u_{iN})$ corresponding to λ_i, and using the spectral decomposition, we have

$$\max_a \mathsf{Im} G_{aa}(z) = \max_a \sum_{i=1}^N \frac{|u_{ia}|^2\eta}{|\lambda_i - E|^2 + \eta^2} = \sum_{i=1}^N \frac{||\mathbf{u}_i||_\infty^2 \eta}{|\lambda_i - E|^2 + \eta^2}. \tag{2.55}$$

For any $|\lambda_i| \le \sqrt{2} - \kappa$, we set $E = \lambda_i$ on the r.h.s. of (2.55) and use (2.54) to bound the l.h.s. of it. Then we obtain

$$\frac{||\mathbf{u}_i||_\infty^2}{\eta} \prec 1.$$

Choosing $\eta = N^{-1+\varepsilon_2}$ above and using the fact that ε_2 can be arbitrarily small, we can get (1.17). Hence, we completed the proof of Theorem 1.11.

3. Supersymmetric formalism and integral representation for the Green's function

In this section, we will represent $\mathbb{E}|G_{ij}(z)|^{2n}$ for the Gaussian case by a superintegral. The final representation is stated in (3.30). We make the convention here, for any real argument in an integral below, its region of the integral is always $\mathbb{R}$, unless specified otherwise.

3.1. **Gaussian integrals and superbosonization formulas.** Let $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_k)'$ be a vector of complex components, $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_k)'$ be a vector of Grassmann components. In addition, let $\boldsymbol{\phi}^*$ and $\boldsymbol{\psi}^*$ be the conjugate transposes of $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$, respectively. We recall the following well-known formulas for Gaussian integrals.

Proposition 3.1 (Gaussian integrals or Wick's formulas)**.**

(i) *Let* A *be a* $k \times k$ *complex matrix with positive-definite Hermitian part, i.e.* $\mathsf{Re}A > 0$. *Then for any* $\ell \in \mathbb{N}$, *and* $i_1, \ldots, i_\ell, j_1, \ldots, j_\ell \in \{1, \ldots, k\}$, *we have*

$$\int \prod_{a=1}^k \frac{\mathrm{d}\mathsf{Re}\phi_a \mathrm{d}\mathsf{Im}\phi_a}{\pi} \exp\{-\boldsymbol{\phi}^* \mathrm{A} \boldsymbol{\phi}\} \prod_{b=1}^\ell \bar{\phi}_{i_b} \phi_{j_b} = \frac{1}{\det \mathrm{A}} \sum_{\sigma\in\mathbb{P}(\ell)} \prod_{b=1}^\ell (\mathrm{A}^{-1})_{j_b, i_{\sigma(b)}}, \tag{3.1}$$

where $\mathbb{P}(\ell)$ *is the permutation group of degree* ℓ.

(ii) *Let* B *be any* $k \times k$ *matrix. Then for any* $\ell \in \{0, \ldots, k\}$, *any* ℓ *distinct integers* $i_1, \ldots, i_\ell$ *and another* ℓ *distinct integers* $j_1, \ldots, j_\ell \in \{1, \ldots, k\}$, *we have*

$$\int \prod_{a=1}^k \mathrm{d}\bar{\psi}_a \mathrm{d}\psi_a \exp\{-\boldsymbol{\psi}^* \mathrm{B} \boldsymbol{\psi}\} \prod_{b=1}^\ell \bar{\psi}_{i_b} \psi_{j_b} = (-1)^{\ell + \sum_{\alpha=1}^\ell (i_\alpha + j_\alpha)} \det \mathrm{B}^{(\mathsf{I}|\mathsf{J})}, \tag{3.2}$$

where $\mathsf{I} = \{i_1, \ldots, i_\ell\}$, *and* $\mathsf{J} = \{j_1, \ldots, j_\ell\}$.

Now, we introduce the superbosonization formula for superintegrals. Let $\boldsymbol{\chi} = (\chi_{ij})$ be an $\ell \times r$ matrix with Grassmann entries, $\mathbf{f} = (f_{ij})$ be an $\ell \times r$ matrix with complex entries. In addition, we denote their conjugate transposes by $\boldsymbol{\chi}^*$ and $\mathbf{f}^*$ respectively. Let F be a function of the entries of the matrix

$$\mathcal{S}(\mathbf{f}, \mathbf{f}^*; \boldsymbol{\chi}, \boldsymbol{\chi}^*) := \begin{pmatrix} \boldsymbol{\chi}^*\boldsymbol{\chi} & \boldsymbol{\chi}^*\mathbf{f} \\ \mathbf{f}^*\boldsymbol{\chi} & \mathbf{f}^*\mathbf{f} \end{pmatrix}.$$

Let $\mathcal{A}(\boldsymbol{\chi}, \boldsymbol{\chi}^*)$ be the Grassmann algebra generated by χ_{ij}'s and $\bar{\chi}_{ij}$'s. Then we can regard F as a function defined on a complex vector space, taking values in $\mathcal{A}(\boldsymbol{\chi}, \boldsymbol{\chi}^*)$. Hence, we can and do view $F(\mathcal{S}(\mathbf{f}, \mathbf{f}^*; \boldsymbol{\chi}, \boldsymbol{\chi}^*))$ as a polynomial in χ_{ij}'s and $\bar{\chi}_{ij}$'s, in which the coefficients are functions of f_{ij}'s and $\bar{f}_{ij}$'s. Under this viewpoint, we state the assumption on F as follows.

Assumption 3.2. *Suppose that $F(\mathcal{S}(\mathbf{f}, \mathbf{f}^*; \boldsymbol{\chi}, \boldsymbol{\chi}^*))$ is a holomorphic function of f_{ij}'s and $\bar{f}_{ij}$'s if they are regarded as independent variables, and F is a Schwarz function of $\mathsf{Re} f_{ij}$'s and $\mathsf{Im} f_{ij}$'s, by those we mean that all of the coefficients of $F(\mathcal{S}(\mathbf{f}, \mathbf{f}^*; \boldsymbol{\chi}, \boldsymbol{\chi}^*))$, as functions of f_{ij}'s and $\bar{f}_{ij}$'s, possess the above properties.*

Proposition 3.3 (Superbosonization formula for the nonsingular case,[17])**.** *Suppose that F satisfies Assumption 3.2. For $\ell \geq r$, we have*

$$\int F\begin{pmatrix} \boldsymbol{\chi}^*\boldsymbol{\chi} & \boldsymbol{\chi}^*\mathbf{f} \\ \mathbf{f}^*\boldsymbol{\chi} & \mathbf{f}^*\mathbf{f} \end{pmatrix} \mathrm{d}\mathbf{f}\mathrm{d}\boldsymbol{\chi} = (\mathbf{i}\pi)^{-r(r-1)} \int \mathrm{d}\hat{\mu}(\mathbf{x})\mathrm{d}\hat{\nu}(\mathbf{y})\mathrm{d}\boldsymbol{\omega}\mathrm{d}\boldsymbol{\xi}\, F\begin{pmatrix} \mathbf{x} & \boldsymbol{\omega} \\ \boldsymbol{\xi} & \mathbf{y} \end{pmatrix} \frac{\det{}^{\ell}\mathbf{y}}{\det{}^{\ell}(\mathbf{x} - \boldsymbol{\omega}\mathbf{y}^{-1}\boldsymbol{\xi})}, \tag{3.3}$$

where $\mathbf{x} = (x_{ij})$ is a unitary matrix; $\mathbf{y} = (y_{ij})$ is a positive-definite Hermitian matrix; $\boldsymbol{\omega}$ and $\boldsymbol{\xi}$ are two Grassmann matrices, and all of them are $r \times r$. Here

$$\mathrm{d}\mathbf{f} = \prod_{i,j} \frac{\mathrm{d}\mathsf{Re} f_{ij}\mathrm{d}\mathsf{Im} f_{ij}}{\pi}, \qquad \mathrm{d}\boldsymbol{\chi} = \prod_{i,j} \mathrm{d}\bar{\chi}_{ij}\mathrm{d}\chi_{ij},$$

$$\mathrm{d}\hat{\nu}(\mathbf{y}) = \mathbf{1}(\mathbf{y} > 0) \prod_{i=1}^{r} \mathrm{d}y_{ii} \prod_{j>k} \mathrm{d}\mathsf{Re} y_{jk}\mathrm{d}\mathsf{Im} y_{jk}, \qquad \mathrm{d}\boldsymbol{\omega}\mathrm{d}\boldsymbol{\xi} = \prod_{i,j=1}^{r} \mathrm{d}\omega_{ij}\mathrm{d}\xi_{ij},$$

and $\mathrm{d}\hat{\mu}(\cdot)$ is defined by

$$\mathrm{d}\hat{\mu}(\mathbf{x}) = \frac{\pi^{r(r-1)/2}}{\prod_{i=1}^{r} i!} \cdot \prod_{i=1}^{r} \frac{\mathrm{d}x_i}{2\pi\mathbf{i}} \cdot (\Delta(x_1, \ldots, x_r))^2 \cdot \mathrm{d}\mu(V),$$

under the parametrization induced by the eigendecomposition, namely,

$$\mathbf{x} = V^*\hat{\mathbf{x}}V, \quad \hat{\mathbf{x}} = \mathrm{diag}(x_1, \ldots, x_r), \quad V \in \mathring{U}(r).$$

Here $\mathrm{d}\mu(V)$ is the Haar measure on $\mathring{U}(r)$, and $\Delta(\cdot)$ is the Vandermonde determinant. In addition, the integral w.r.p.t. $\mathbf{x}$ ranges over $U(2)$, that w.r.p.t. $\mathbf{y}$ ranges over all positive-definite matrices.

For the singular case, i.e. $r > \ell$, we only state the formula for the case of $r = 2$ and $\ell = 1$, which is enough for our purpose. We can refer to formula (11) in [3] for the result under more general setting on r and ℓ.

Proposition 3.4 (Superbosonization formula for the singular case,[3])**.** *Suppose that F satisfies Assumption 3.2. If $r = 2$ and $\ell = 1$, we have*

$$\int F\begin{pmatrix} \boldsymbol{\chi}^*\boldsymbol{\chi} & \boldsymbol{\chi}^*\mathbf{f} \\ \mathbf{f}^*\boldsymbol{\chi} & \mathbf{f}^*\mathbf{f} \end{pmatrix} \mathrm{d}\mathbf{f}\mathrm{d}\boldsymbol{\chi} = \frac{-1}{\pi^2} \int \mathrm{d}\mathbf{w}\mathrm{d}\hat{\mu}(\mathbf{x}) \cdot \mathbf{1}(y \geq 0)\mathrm{d}y \cdot \mathrm{d}\boldsymbol{\omega}\mathrm{d}\boldsymbol{\xi}\, F\begin{pmatrix} \mathbf{x} & \boldsymbol{\omega}\mathbf{w}^* \\ \mathbf{w}\boldsymbol{\xi} & y\mathbf{w}\mathbf{w}^* \end{pmatrix} \frac{y(y - \boldsymbol{\xi}\mathbf{x}^{-1}\boldsymbol{\omega})^2}{\det{}^2\mathbf{x}}, \tag{3.4}$$

where y is a positive variable; $\mathbf{x}$ *is a 2-dimensional unitary matrix;* $\boldsymbol{\omega} = (\omega_1, \omega_2)'$ *and* $\boldsymbol{\xi} = (\xi_1, \xi_2)$ *are two vectors with Grassmann components. In addition,* $\mathbf{w}$ *is a unit vector, which can be parameterized by*

$$\mathbf{w} = \Big(1/\sqrt{1+|w|^2}, w/\sqrt{1+|w|^2}\Big)', \quad w \in \mathbb{C}.$$

Moreover, the differentials are defined as

$$\mathrm{d}\mathbf{w} = \frac{1}{(1+|w|^2)^2}\mathrm{dRe}w\mathrm{dIm}w, \qquad \mathrm{d}\boldsymbol{\omega}\mathrm{d}\boldsymbol{\xi} = \prod_{i=1,2}\mathrm{d}\omega_i\mathrm{d}\xi_i.$$

In addition, the integral w.r.p.t. $\mathbf{x}$ *ranges over* $U(2)$.

In our discussion, for $\mathbf{w}$, we will adopt the parametrization

$$v = 1/\sqrt{1+|w|^2}, \quad e^{\mathbf{i}\theta} = w/|w|, \quad v \in \mathbb{I}, \quad \theta \in \mathbb{L}$$

for convenience. Accordingly, we can get

$$\mathrm{d}\mathbf{w} = v\mathrm{d}v\mathrm{d}\theta.$$

3.2. **Initial representation.** For $a = 1, 2$ and $j = 1, \ldots, W$, we set

$$\Phi_{a,j} = \big(\phi_{a,j,1}, \ldots, \phi_{a,j,M}\big)', \quad \Psi_{a,j} = \big(\psi_{a,j,1}, \ldots, \psi_{a,j,M}\big)'$$
$$\Phi_a = \big(\Phi'_{a,1}, \ldots, \Phi'_{a,W}\big)', \quad \Psi_a = \big(\Psi'_{a,1}, \ldots, \Psi'_{a,W}\big)'.$$

For each j and each a, $\Phi_{a,j}$ is a vector with complex components, and $\Psi_{a,j}$ is a vector with Grassmann components. In addition, we use $\Phi^*_{a,j}$ and $\Psi^*_{a,j}$ to represent the conjugate transposes of $\Phi_{a,j}$ and $\Psi_{a,j}$ respectively. Analogously, we adopt the notation Φ^*_a and Ψ^*_a to represent the conjugate transposes of Φ_a and Ψ_a, respectively. We have the following integral representation for the moments of the Green's function.

Lemma 3.5. *For any* $p, q = 1, \ldots, W$ *and* $\alpha, \beta = 1, \ldots, M$, *we have*

$$|G_{pq,\alpha\beta}(z)|^{2n} = \frac{1}{(n!)^2}\int \mathrm{d}\Phi\mathrm{d}\Psi\ \big(\bar{\phi}_{1,q,\beta}\phi_{1,p,\alpha}\bar{\phi}_{2,p,\alpha}\phi_{2,q,\beta}\big)^n$$
$$\times \exp\Big\{\mathbf{i}\Psi^*_1(z-H)\Psi_1 + \mathbf{i}\Phi^*_1(z-H)\Phi_1 - \mathbf{i}\Psi^*_2(\bar{z}-H)\Psi_2 - \mathbf{i}\Phi^*_2(\bar{z}-H)\Phi_2\Big\}, \tag{3.5}$$

where

$$\mathrm{d}\Phi = \prod_{a=1,2}\prod_{j=1}^{W}\prod_{\alpha'=1}^{M}\frac{\mathrm{dRe}\phi_{a,j,\alpha'}\mathrm{dIm}\phi_{a,j,\alpha'}}{\pi}, \quad \mathrm{d}\Psi = \prod_{a=1,2}\prod_{j=1}^{W}\prod_{\alpha'=1}^{M}\mathrm{d}\bar{\psi}_{a,j,\alpha'}\mathrm{d}\psi_{a,j,\alpha'}.$$

Proof. By using Proposition 3.1 (i) with $\ell = n$ and Proposition 3.1 (ii) with $\ell = 0$, we can get (3.5) immediately. □

3.3. **Averaging over the Gaussian random matrix.** Recall the variance profile $\tilde{S}$ in (1.2). Now, we take expectation of the Green's function, i.e average over the random matrix. By elementary Gaussian integral, we get

$$\mathbb{E}|G_{pq,\alpha\beta}(z)|^{2n} = \frac{1}{(n!)^2}\int \mathrm{d}\Phi\mathrm{d}\Psi\ \big(\bar{\phi}_{1,q,\beta}\phi_{1,p,\alpha}\bar{\phi}_{2,p,\alpha}\phi_{2,q,\beta}\big)^n\ \exp\Big\{\mathbf{i}\sum_{j=1}^{W}(Tr\breve{X}_jJZ + Tr\breve{Y}_jJZ)\Big\}$$
$$\times \exp\Big\{\frac{1}{2M}\sum_{j,k}\tilde{\mathfrak{s}}_{jk}Tr\breve{X}_jJ\breve{X}_kJ - \frac{1}{2M}\sum_{j,k}\tilde{\mathfrak{s}}_{jk}Tr\breve{Y}_jJ\breve{Y}_kJ\Big\}$$
$$\times \exp\Big\{-\frac{1}{M}\sum_{j,k}\tilde{\mathfrak{s}}_{jk}Tr\breve{\Omega}_jJ\breve{\Xi}_kJ\Big\}, \tag{3.6}$$

where

$$J = \mathrm{diag}(1, -1), \quad Z = \mathrm{diag}(z, \bar{z}),$$

and for each $j = 1, \ldots, W$, the matrices $\breve{X}_j$, $\breve{Y}_j$, $\breve{\Omega}_j$ and $\breve{\Xi}_j$ are 2×2 blocks of a supermatrix, namely,

$$\breve{\mathcal{S}}_j = \left(\begin{array}{c|c} \breve{X}_j & \breve{\Omega}_j \\ \hline \breve{\Xi}_j & \breve{Y}_j \end{array}\right) := \left(\begin{array}{cc|cc} \Psi_{1,j}^*\Psi_{1,j} & \Psi_{1,j}^*\Psi_{2,j} & \Psi_{1,j}^*\Phi_{1,j} & \Psi_{1,j}^*\Phi_{2,j} \\ \Psi_{2,j}^*\Psi_{1,j} & \Psi_{2,j}^*\Psi_{2,j} & \Psi_{2,j}^*\Phi_{1,j} & \Psi_{2,j}^*\Phi_{2,j} \\ \hline \Phi_{1,j}^*\Psi_{1,j} & \Phi_{1,j}^*\Psi_{2,j} & \Phi_{1,j}^*\Phi_{1,j} & \Phi_{1,j}^*\Phi_{2,j} \\ \Phi_{2,j}^*\Psi_{1,j} & \Phi_{2,j}^*\Psi_{2,j} & \Phi_{2,j}^*\Phi_{1,j} & \Phi_{2,j}^*\Phi_{2,j} \end{array}\right).$$

Remark 3.6. *The derivation of (3.6) from (3.5) is quite standard. We refer to the proof of (2.14) in [19] for more details and will not reproduce it here.*

3.4. **Decomposition of the supermatrices.** From now on, we split the discussion into the following three cases

- (Case 1): Entries in the off-diagonal blocks, i.e. $p \neq q$,
- (Case 2): Off-diagonal entries in the diagonal blocks, i.e. $p = q, \quad \alpha \neq \beta$,
- (Case 3): Diagonal entries, i.e. $p = q, \quad \alpha = \beta$.

For each case, we will perform a decomposition for the supermatrix $\breve{\mathcal{S}}_j$ ($j = p$ or q). For a vector $\mathbf{v}$ and some index set I, we use $\mathbf{v}^{\langle \mathsf{I} \rangle}$ to denote the subvector obtained by deleting the i-th component of $\mathbf{v}$ for all $i \in \mathsf{I}$. Then, we adopt the notation

$$\breve{\mathcal{S}}_j^{\langle \mathsf{I} \rangle} = \left(\begin{array}{c|c} \breve{X}_j^{\langle \mathsf{I} \rangle} & \breve{\Omega}_j^{\langle \mathsf{I} \rangle} \\ \hline \breve{\Xi}_j^{\langle \mathsf{I} \rangle} & \breve{Y}_j^{\langle \mathsf{I} \rangle} \end{array}\right), \qquad \breve{\mathcal{S}}_j^{[i]} = \left(\begin{array}{c|c} \breve{X}_j^{[i]} & \breve{\Omega}_j^{[i]} \\ \hline \breve{\Xi}_j^{[i]} & \breve{Y}_j^{[i]} \end{array}\right).$$

Here, for $\mathsf{A} = \breve{X}_j$, $\breve{Y}_j$, $\breve{\Omega}_j$ or $\breve{\Xi}_j$, the notation $\mathsf{A}^{\langle \mathsf{I} \rangle}$ is defined via replacing $\Phi_{a,j}$, $\Psi_{a,j}$, $\Phi_{a,j}^*$ and $\Psi_{a,j}^*$ by $\Phi_{a,j}^{\langle \mathsf{I} \rangle}$, $\Psi_{a,j}^{\langle \mathsf{I} \rangle}$, $(\Phi_{a,j}^*)^{\langle \mathsf{I} \rangle}$ and $(\Psi_{a,j}^*)^{\langle \mathsf{I} \rangle}$, respectively, for $a = 1, 2$, in the definition of A. In addition, the notation $\mathsf{A}^{[i]}$ is defined via replacing $\Phi_{a,j}$, $\Psi_{a,j}$, $\Phi_{a,j}^*$ and $\Psi_{a,j}^*$ by $\phi_{a,j,i}$, $\psi_{a,j,i}$, $\bar{\phi}_{a,j,i}$ and $\bar{\psi}_{a,j,i}$ respectively, for $a = 1, 2$, in the definition of A. Moreover, for $\mathsf{A} = \breve{\mathcal{S}}_j$, $\breve{X}_j$, $\breve{Y}_j$, $\breve{\Omega}_j$ or $\breve{\Xi}_j$, we will simply abbreviate $\mathsf{A}^{\langle \{a,b\} \rangle}$ and $\mathsf{A}^{\langle \{a\} \rangle}$ by $\mathsf{A}^{\langle a,b \rangle}$ and $\mathsf{A}^{\langle a \rangle}$, respectively. Note that $\breve{\mathcal{S}}_j^{[i]}$ is of rank-one.

For Case 1, due to symmetry, we can assume $\alpha = \beta = 1$. Then we extract two rank-one supermatrices from $\breve{\mathcal{S}}_p$ and $\breve{\mathcal{S}}_q$ such that the quantities $\bar{\phi}_{2,p,1}\phi_{1,p,1}$ and $\bar{\phi}_{1,q,1}\phi_{2,q,1}$ can be expressed in terms of the entries of these supermatrices. More specifically, we decompose the supermatrices

$$\breve{\mathcal{S}}_p = \breve{\mathcal{S}}_p^{\langle 1 \rangle} + \breve{\mathcal{S}}_p^{[1]}, \quad \breve{\mathcal{S}}_q = \breve{\mathcal{S}}_q^{\langle 1 \rangle} + \breve{\mathcal{S}}_q^{[1]}. \tag{3.7}$$

Consequently, we can write

$$\bar{\phi}_{1,q,1}\phi_{1,p,1}\bar{\phi}_{2,p,1}\phi_{2,q,1} = (\breve{Y}_q^{[1]})_{12}(\breve{Y}_p^{[1]})_{21}. \tag{3.8}$$

For Case 2, due to symmetry, we can assume that $\alpha = 1$, $\beta = 2$. Then we extract two rank-one supermatrices from $\breve{\mathcal{S}}_p$, namely,

$$\breve{\mathcal{S}}_p = \breve{\mathcal{S}}_p^{\langle 1,2 \rangle} + \breve{\mathcal{S}}_p^{[1]} + \breve{\mathcal{S}}_p^{[2]}. \tag{3.9}$$

Consequently, we can write

$$\bar{\phi}_{1,p,2}\phi_{1,p,1}\bar{\phi}_{2,p,1}\phi_{2,p,2} = (\breve{Y}_p^{[2]})_{12}(\breve{Y}_p^{[1]})_{21}. \tag{3.10}$$

Finally, for Case 3, due to symmetry, we can assume that $\alpha = 1$. Then we extract only one rank-one supermatrix from $\breve{\mathcal{S}}_p$, namely,

$$\breve{\mathcal{S}}_p = \breve{\mathcal{S}}_p^{\langle 1 \rangle} + \breve{\mathcal{S}}_p^{[1]}. \tag{3.11}$$

Consequently, we can write

$$\bar{\phi}_{1,p,1}\phi_{1,p,1}\bar{\phi}_{2,p,1}\phi_{2,p,1} = (\check{Y}_p^{[1]})_{12}(\check{Y}_p^{[1]})_{21} = (\check{Y}_p^{[1]})_{11}(\check{Y}_p^{[1]})_{22}.$$

Since the discussion for all three cases are similar, we will only present the details for Case 1. More specifically, in the remaining part of this section and Section 4 to Section 10, we will only treat Case 1. In Section 11, we will sum up the discussions in the previous sections and explain how to adapt them to Case 2 and Case 3, resulting a final proof of Theorem 1.14.

3.5. **Variable reduction by superbosonization formulae.** We will work with Case 1. Recall the decomposition (3.7). We use the superbosonization formulae to reduce the number of variables. We shall treat $\check{\mathcal{S}}_k$ $(k \neq p, q)$ and $\check{\mathcal{S}}_j^{\langle 1\rangle}$ $(j = p, q)$ on an equal footing and use the formula (3.3) with $r = 2, \ell = M$ for the former and $r = 2, \ell = M - 1$ for the latter, while we separate the terms $\check{\mathcal{S}}_j^{[1]}$ $(j = p, q)$ and use the formula (3.4). For simplicity, we introduce the notation

$$\tilde{\mathcal{S}}_j = \begin{cases} \check{\mathcal{S}}_j, & \text{if} \quad j \neq p, q, \\ \\ \check{\mathcal{S}}_j^{\langle 1\rangle}, & \text{if} \quad j = p, q. \end{cases} \tag{3.12}$$

Accordingly, we will use $\tilde{X}_j$, $\tilde{\Omega}_j$, $\tilde{\Xi}_j$ and $\tilde{Y}_j$ to denote four blocks of $\tilde{\mathcal{S}}_j$. With this notation, we can rewrite (3.6) with $\alpha = \beta = 1$ as

$$\begin{aligned}
\mathbb{E}|G_{pq,11}(z)|^{2n} = {} & \frac{1}{(n!)^2}\int \mathrm{d}\Phi\mathrm{d}\Psi\ \Big(\bar{\phi}_{1,q,1}\phi_{1,p,1}\bar{\phi}_{2,p,1}\phi_{2,q,1}\Big)^n\ \exp\Big\{\mathbf{i}\sum_{j=1}^{W}\Big(Tr\tilde{X}_jJZ + Tr\tilde{Y}_jJZ\Big)\Big\} \\
& \times \exp\Big\{\frac{1}{2M}\sum_{j,k}\tilde{\mathfrak{s}}_{jk}\Big(Tr\tilde{X}_jJ\tilde{X}_kJ - Tr\tilde{Y}_jJ\tilde{Y}_kJ\Big)\Big\}\ \exp\Big\{-\frac{1}{M}\sum_{j,k}\tilde{\mathfrak{s}}_{jk}Tr\tilde{\Omega}_jJ\tilde{\Xi}_kJ\Big\} \\
& \times \prod_{k=p,q}\exp\Big\{\mathbf{i}Tr\check{X}_k^{[1]}JZ + \mathbf{i}Tr\check{Y}_k^{[1]}JZ\Big\}\ \prod_{k=p,q}\exp\Big\{\frac{1}{M}\sum_{j=1}^{W}\tilde{\mathfrak{s}}_{jk}\Big(Tr\tilde{X}_jJ\check{X}_k^{[1]}J - Tr\tilde{Y}_jJ\check{Y}_k^{[1]}J\Big)\Big\} \\
& \times \prod_{k,\ell=p,q}\exp\Big\{\frac{\tilde{\mathfrak{s}}_{k\ell}}{2M}\Big(Tr\check{X}_k^{[1]}J\check{X}_\ell^{[1]}J - Tr\check{Y}_k^{[1]}J\check{Y}_\ell^{[1]}J\Big)\Big\}\ \prod_{k,\ell=p,q}\exp\Big\{-\frac{\tilde{\mathfrak{s}}_{k\ell}}{M}Tr\check{\Omega}_k^{[1]}J\check{\Xi}_\ell^{[1]}J\Big\} \\
& \times \prod_{k=p,q}\exp\Big\{-\frac{1}{M}\sum_{j}\tilde{\mathfrak{s}}_{jk}\Big(Tr\tilde{\Omega}_jJ\check{\Xi}_k^{[1]}J + Tr\check{\Omega}_k^{[1]}J\tilde{\Xi}_jJ\Big)\Big\}
\end{aligned} \tag{3.13}$$

Now, we use the superbosonization formulae, i.e., (3.3) and (3.4), to change to the reduced variables as

$$\begin{aligned}
&\tilde{X}_j \to X_j, \quad \tilde{Y}_j \to Y_j, \quad \tilde{\Omega}_j \to \Omega_j, \quad \tilde{\Xi}_j \to \Xi_j, \quad j = 1, \ldots, W, \\
&\check{X}_k^{[1]} \to X_k^{[1]}, \quad \check{\Omega}_k^{[1]} \to \boldsymbol{\omega}_k^{[1]}(\mathbf{w}_k^{[1]})^*, \quad \check{\Xi}_k^{[1]} \to \mathbf{w}_k^{[1]}\boldsymbol{\xi}_k^{[1]}, \quad \check{Y}_k^{[1]} \to Y_k^{[1]} := y_k^{[1]}\mathbf{w}_k^{[1]}(\mathbf{w}_k^{[1]})^*, \quad k = p, q.
\end{aligned} \tag{3.14}$$

Here, for $j = 1, \ldots, W$, X_j is a 2×2 unitary matrix; Y_j is a 2×2 positive-definite matrix; $\Omega_j = (\omega_{j,\alpha\beta})$ and $\Xi_j = (\xi_{j,\alpha\beta})$ are 2×2 Grassmann matrices. For $k = p$ or q, $X_k^{[1]}$ is a 2×2 unitary matrix; $y_k^{[1]}$ is a positive variable; $\boldsymbol{\omega}_k^{[1]} = (\omega_{k,1}^{[1]}, \omega_{k,2}^{[1]})'$ is a column vector with Grassmann components; $\boldsymbol{\xi}_k^{[1]} = (\xi_{k,1}^{[1]}, \xi_{k,2}^{[1]})$ is a row vector with Grassmann components. In addition, for $k = p, q$,

$$\mathbf{w}_k^{[1]} = \Big(v_k^{[1]}, u_k^{[1]}e^{\mathbf{i}\sigma_k^{[1]}}\Big)', \quad u_k^{[1]} = \sqrt{1 - (v_k^{[1]})^2}, \quad v_k^{[1]} \in \mathbb{I}, \quad \sigma_k^{[1]} \in \mathbb{L}. \tag{3.15}$$

Then by using superbosonization formulae, we arrive at the representation

$$\mathbb{E}|G_{pq,11}(z)|^{2n}$$

$$= \frac{(-1)^W}{(n!)^2\pi^{2W+4}} \int \mathrm{d}X^{[1]}\mathrm{d}\mathbf{y}^{[1]}\mathrm{d}\mathbf{w}^{[1]}\mathrm{d}\boldsymbol{\omega}^{[1]}\mathrm{d}\boldsymbol{\xi}^{[1]}\mathrm{d}X\mathrm{d}Y\mathrm{d}\Omega\mathrm{d}\Xi \left(y_p^{[1]}y_q^{[1]}\big(\mathbf{w}_q^{[1]}(\mathbf{w}_q^{[1]})^*\big)_{12}\big(\mathbf{w}_p^{[1]}(\mathbf{w}_p^{[1]})^*\big)_{21}\right)^n$$
$$\times \exp\Big\{\mathbf{i}\sum_{j=1}^W \Big(TrX_jJZ + TrY_jJZ\Big)\Big\} \exp\Big\{\frac{1}{2M}\sum_{j,k}\tilde{\mathfrak{s}}_{jk}\Big(TrX_jJX_kJ - TrY_jJY_kJ\Big)\Big\}$$
$$\times \exp\Big\{-\frac{1}{M}\sum_{j,k}\tilde{\mathfrak{s}}_{jk}Tr\Omega_jJ\Xi_kJ\Big\} \prod_j \frac{\det{}^M Y_j}{\det{}^M\big(X_j - \Omega_jY_j^{-1}\Xi_j\big)} \prod_{k=p,q}\frac{\det\big(X_k - \Omega_kY_k^{-1}\Xi_k\big)}{\det Y_k}$$
$$\times \prod_{k=p,q}\exp\Big\{\mathbf{i}TrX_k^{[1]}JZ + \mathbf{i}TrY_k^{[1]}JZ\Big\} \prod_{k=p,q}\exp\Big\{\frac{1}{M}\sum_{j=1}^W\tilde{\mathfrak{s}}_{jk}\Big(TrX_jJX_k^{[1]}J - TrY_jJY_k^{[1]}J\Big)\Big\}$$
$$\times \prod_{k,\ell=p,q}\exp\Big\{\frac{\tilde{\mathfrak{s}}_{k\ell}}{2M}\Big(TrX_k^{[1]}JX_\ell^{[1]}J - TrY_k^{[1]}JY_\ell^{[1]}J\Big)\Big\} \prod_{k,\ell=p,q}\exp\Big\{-\frac{\tilde{\mathfrak{s}}_{k\ell}}{M}Tr\boldsymbol{\omega}_k^{[1]}(\mathbf{w}_k^{[1]})^*J\mathbf{w}_\ell^{[1]}\boldsymbol{\xi}_\ell^{[1]}J\Big\}$$
$$\times \prod_{k=p,q}\exp\Big\{-\frac{1}{M}\sum_{j=1}^W\tilde{\mathfrak{s}}_{jk}Tr\Omega_jJ\mathbf{w}_k^{[1]}\boldsymbol{\xi}_k^{[1]}J\Big\} \prod_{k=p,q}\exp\Big\{-\frac{1}{M}\sum_{j=1}^W\tilde{\mathfrak{s}}_{jk}Tr\boldsymbol{\omega}_k^{[1]}(\mathbf{w}_k^{[1]})^*J\Xi_jJ\Big\}$$
$$\times \prod_{k=p,q}\frac{y_k^{[1]}\Big(y_k^{[1]} - \boldsymbol{\xi}_k^{[1]}(X_k^{[1]})^{-1}\boldsymbol{\omega}_k^{[1]}\Big)^2}{\det{}^2(X_k^{[1]})}, \tag{3.16}$$

where we used the notation $\mathbf{y}^{[1]} := (y_p^{[1]}, y_q^{[1]})$, $\mathbf{w}^{[1]} := (\mathbf{w}_p^{[1]}, \mathbf{w}_q^{[1]})$. The differentials in (3.16) are defined by

$$\mathrm{d}X^{[1]} := \prod_{j=p,q}\mathrm{d}\hat{\mu}(X_p^{[1]})\mathrm{d}\hat{\mu}(X_q^{[1]}), \qquad \mathrm{d}\mathbf{y}^{[1]} := \prod_{j=p,q}\mathbf{1}(y_j^{[1]}>0)\mathrm{d}y_j^{[1]},$$
$$\mathrm{d}\mathbf{w}^{[1]} := \prod_{j=p,q}\mathrm{d}\mathbf{w}_j^{[1]} = \prod_{j=p,q}v_j^{[1]}\mathrm{d}v_j^{[1]}\mathrm{d}\sigma_j^{[1]}, \qquad \mathrm{d}\boldsymbol{\omega}^{[1]}\mathrm{d}\boldsymbol{\xi}^{[1]} := \prod_{\alpha=1,2}\prod_{j=p,q}\omega_{j,\alpha}^{[1]}\xi_{j,\alpha}^{[1]}.$$
$$\mathrm{d}X := \prod_{j=1}^W\mathrm{d}\hat{\mu}(X_j), \qquad \mathrm{d}Y := \prod_{j=1}^W\mathrm{d}\hat{\nu}(Y_j), \qquad \mathrm{d}\Omega\mathrm{d}\Xi := \prod_{\alpha,\beta=1,2}\prod_{j=1}^W\mathrm{d}\omega_{j,\alpha\beta}\mathrm{d}\xi_{j,\alpha\beta}.$$

The regions of the integral of X_j's and $X_k^{[1]}$'s are all $U(2)$, and those of Y_j's are the set of all positive-definite matrices. The integral of $v_k^{[1]}$ ranges over $\mathbb{I}$ and that of $\sigma_k^{[1]}$ ranges over $\mathbb{L}$, for $k = p, q$.

Now we change the variables as $X_jJ \to X_j, Y_jJ \to B_j, \Omega_jJ \to \Omega_j, \Xi_jJ \to \Xi_j$ and perform the scaling $X_j \to -MX_j$, $B_j \to MB_j$, $\Omega_j \to \sqrt{M}\Omega_j$ and $\Xi_j \to \sqrt{M}\Xi_j$. Consequently, we can write

$$\mathbb{E}|G_{pq,11}(z)|^{2n} = \frac{(-1)^WM^{4W}}{(n!)^2\pi^{2W+4}}\int \mathrm{d}X^{[1]}\mathrm{d}\mathbf{y}^{[1]}\mathrm{d}\mathbf{w}^{[1]}\mathrm{d}\boldsymbol{\omega}^{[1]}\mathrm{d}\boldsymbol{\xi}^{[1]}\mathrm{d}X\mathrm{d}B\mathrm{d}\Omega\mathrm{d}\Xi \exp\Big\{-M\big(K(X)+L(B)\big)\Big\}$$
$$\times\mathcal{P}(\Omega,\Xi,X,B)\cdot\mathcal{Q}\big(\Omega,\Xi,\boldsymbol{\omega}^{[1]},\boldsymbol{\xi}^{[1]},X^{[1]},\mathbf{y}^{[1]},\mathbf{w}^{[1]}\big)\cdot\mathcal{F}\big(X,B,X^{[1]},\mathbf{y}^{[1]},\mathbf{w}^{[1]}\big), \tag{3.17}$$

where the functions in the integrand are defined as

$$K(X) := -\frac{1}{2}\sum_{j,k}\tilde{\mathfrak{s}}_{jk}TrX_jX_k + \mathbf{i}E\sum_j TrX_j + \sum_j\log\det X_j,$$
$$L(B) := \frac{1}{2}\sum_{j,k}\tilde{\mathfrak{s}}_{jk}TrB_jB_k - \mathbf{i}E\sum_j TrB_j - \sum_j\log\det B_j,$$
$$\mathcal{P}(\Omega,\Xi,X,B) := \exp\Big\{-\sum_{j,k}\tilde{\mathfrak{s}}_{jk}Tr\Omega_j\Xi_k\Big\}\prod_j\frac{1}{\det{}^M\big(1+M^{-1}X_j^{-1}\Omega_jB_j^{-1}\Xi_j\big)}$$
$$\times\prod_{k=p,q}\frac{\det\big(X_k+M^{-1}\Omega_kB_k^{-1}\Xi_k\big)}{\det B_k},$$

$$\mathcal{Q}(\Omega,\Xi,\boldsymbol{\omega}^{[1]},\boldsymbol{\xi}^{[1]},X^{[1]},\mathbf{y}^{[1]},\mathbf{w}^{[1]}):=\prod_{k=p,q}\Big(1-(y_k^{[1]})^{-1}\boldsymbol{\xi}_k^{[1]}(X_k^{[1]})^{-1}\boldsymbol{\omega}_k^{[1]}\Big)^2$$
$$\times\prod_{k=p,q}\exp\Big\{-\frac{1}{\sqrt{M}}\sum_j\tilde{\mathfrak{s}}_{jk}\Big(Tr\Omega_j\mathbf{w}_k^{[1]}\boldsymbol{\xi}_k^{[1]}J+Tr\boldsymbol{\omega}_k^{[1]}(\mathbf{w}_k^{[1]})^*J\Xi_j\Big)\Big\}$$
$$\times\prod_{k,\ell=p,q}\exp\Big\{-\frac{1}{M}\tilde{\mathfrak{s}}_{k\ell}Tr\boldsymbol{\omega}_k^{[1]}(\mathbf{w}_k^{[1]})^*J\mathbf{w}_\ell^{[1]}\boldsymbol{\xi}_\ell^{[1]}J\Big\},$$

$$\mathcal{F}(X,B,X^{[1]},\mathbf{y}^{[1]},\mathbf{w}^{[1]}):=f(X,X^{[1]})\,g(B,\mathbf{y}^{[1]},\mathbf{w}^{[1]}),\tag{3.18}$$

with

$$f(X,X^{[1]}):=\exp\Big\{M\eta\sum_{j=1}^W TrX_jJ\Big\}\prod_{k,\ell=p,q}\exp\Big\{\frac{\tilde{\mathfrak{s}}_{k\ell}}{2M}TrX_k^{[1]}JX_\ell^{[1]}J\Big\}$$
$$\times\prod_{k=p,q}\frac{1}{\det{}^2(X_k^{[1]})}\exp\Big\{\mathbf{i}TrX_k^{[1]}JZ-\sum_j\tilde{\mathfrak{s}}_{jk}TrX_jX_k^{[1]}J\Big\},\tag{3.19}$$

$$g(B,\mathbf{y}^{[1]},\mathbf{w}^{[1]}):=\exp\Big\{-M\eta\sum_{j=1}^W TrB_jJ\Big\}\prod_{k,\ell=p,q}\exp\Big\{-\frac{\tilde{\mathfrak{s}}_{k\ell}}{2M}TrY_k^{[1]}JY_\ell^{[1]}J\Big\}$$
$$\times\Big((\mathbf{w}_q^{[1]}(\mathbf{w}_q^{[1]})^*)_{12}(\mathbf{w}_p^{[1]}(\mathbf{w}_p^{[1]})^*)_{21}\Big)^n\prod_{k=p,q}(y_k^{[1]})^{n+3}\exp\Big\{\mathbf{i}TrY_k^{[1]}JZ-\sum_j\tilde{\mathfrak{s}}_{jk}TrB_jY_k^{[1]}J\Big\}.\tag{3.20}$$

In (3.17), the regions of X_j's and $X_k^{[1]}$'s are all $U(2)$, and those of B_j's are the set of the matrices A satisfying $AJ>0$. Roughly speaking, here we collected the terms containing the Grassmann variables from $\boldsymbol{\omega}_k^{[1]}$'s and $\boldsymbol{\xi}_k^{[1]}$'s, resulting the factor $\mathcal{Q}(\Omega,\Xi,\boldsymbol{\omega}^{[1]},\boldsymbol{\xi}^{[1]},X^{[1]},\mathbf{y}^{[1]},\mathbf{w}^{[1]})$. Then, we put the terms containing the Grassmann variables in Ω_j's and Ξ_j's together (except those in $\mathcal{Q}(\Omega,\Xi,\boldsymbol{\omega}^{[1]},\boldsymbol{\xi}^{[1]},X^{[1]},\mathbf{y}^{[1]},\mathbf{w}^{[1]})$), resulting the factor $\mathcal{P}(\Omega,\Xi,X,B)$. Finally, we sorted out the terms containing η, $y_k^{[1]}$ and the variables from $X_k^{[1]}$'s and $\mathbf{w}_k^{[1]}$'s (except those in $\mathcal{Q}(\Omega,\Xi,\boldsymbol{\omega}^{[1]},\boldsymbol{\xi}^{[1]},X^{[1]},\mathbf{y}^{[1]},\mathbf{w}^{[1]})$), resulting the factor $\mathcal{F}(X,B,X^{[1]},\mathbf{y}^{[1]},\mathbf{w}^{[1]})$. This separation indicates the order in which we will perform the integrations.

3.6. **Parametrization for X, B.** Similarly to the discussion in [19], we start with some preliminary parameterization. At first, we do the eigendecomposition

$$X_j=P_j^*\hat{X}_jP_j,\quad B_j=Q_j^{-1}\hat{B}_jQ_j,\qquad P_j\in\mathring{U}(2),\quad Q_j\in\mathring{U}(1,1),\tag{3.21}$$

where

$$\hat{X}_j=\mathrm{diag}(x_{j,1},x_{j,2}),\quad \hat{B}_j=\mathrm{diag}(b_{j,1},-b_{j,2}),\quad x_{j,1},x_{j,2}\in\Sigma,\quad b_{j,1},b_{j,2}\in\mathbb{R}_+.\tag{3.22}$$

Further, we introduce

$$V_j=P_jP_1^*\in\mathring{U}(2),\quad T_j=Q_jQ_1^{-1}\in\mathring{U}(1,1),\quad j=1,\dots,W.\tag{3.23}$$

Especially, we have $V_1=T_1=I$. Now, we parameterize P_1, Q_1, V_j and T_j for all $j=2,\dots,W$ as follows

$$P_1=\begin{pmatrix}u & ve^{\mathbf{i}\theta}\\ -ve^{-\mathbf{i}\theta} & u\end{pmatrix},\quad V_j=\begin{pmatrix}u_j & v_je^{\mathbf{i}\theta_j}\\ -v_je^{-\mathbf{i}\theta_j} & u_j\end{pmatrix},$$

$$u=\sqrt{1-v^2},\quad u_j=\sqrt{1-v_j^2},\quad v,v_j\in\mathbb{I},\quad \theta,\theta_j\in\mathbb{L},$$

$$Q_1=\begin{pmatrix}s & te^{\mathbf{i}\sigma}\\ te^{-\mathbf{i}\sigma} & s\end{pmatrix},\quad T_j=\begin{pmatrix}s_j & t_je^{\mathbf{i}\sigma_j}\\ t_je^{-\mathbf{i}\sigma_j} & s_j\end{pmatrix},$$

$$s=\sqrt{1+t^2},\quad s_j=\sqrt{1+t_j^2},\quad t,t_j\in\mathbb{R}_+,\quad \sigma,\sigma_j\in\mathbb{L}. \tag{3.24}$$

Under the parametrization above, we can express the corresponding differentials as follows.

$$\begin{aligned}\mathrm{d}X\mathrm{d}B=\mathrm{d}\mu(P_1)\mathrm{d}\nu(Q_1)\cdot\prod_{j=2}^{W}\mathrm{d}\mu(V_j)\mathrm{d}\nu(T_j)\cdot\prod_{j=1}^{W}\mathrm{d}b_{j,1}\mathrm{d}b_{j,2}\cdot\frac{\mathrm{d}x_{j,1}}{2\pi\mathbf{i}}\frac{\mathrm{d}x_{j,2}}{2\pi\mathbf{i}}\\ \times 2^W(\pi/2)^{2W}\prod_{j=1}^{W}(x_{j,1}-x_{j,2})^2(b_{j,1}+b_{j,2})^2,\end{aligned} \tag{3.25}$$

where

$$\mathrm{d}\mu(P_1)=2v\mathrm{d}v\cdot\frac{\mathrm{d}\theta}{2\pi},\quad \mathrm{d}\mu(V_j)=2v_j\mathrm{d}v_j\cdot\frac{\mathrm{d}\theta_j}{2\pi},\quad \mathrm{d}\nu(Q_1)=2t\mathrm{d}t\cdot\frac{\mathrm{d}\sigma}{2\pi},\quad \mathrm{d}\nu(T_j)=2t_j\mathrm{d}t_j\cdot\frac{\mathrm{d}\sigma_j}{2\pi}.$$

In addition, for simplicity, we do the change of variables

$$\Omega_j\to P_1^*\Omega_jQ_1,\quad \Xi_j\to Q_1^{-1}\Xi_jP_1. \tag{3.26}$$

Note that the Berezinian of such a change is 1. After this change, $\mathcal{P}(\Omega,\Xi,X,B,\mathbf{y}^{[1]},\mathbf{w}^{[1]})$ turns out to be independent of P_1 and Q_1.

To adapt to the new parametrization, we change the notation

$$\begin{aligned}&K(X)\to K(\hat X,V),\quad L(B)\to L(\hat B,T),\quad \mathcal{P}(\Omega,\Xi,X,B)\to\mathcal{P}(\Omega,\Xi,\hat X,\hat B,V,T),\\ &\mathcal{Q}(\Omega,\Xi,\boldsymbol{\omega}^{[1]},\boldsymbol{\xi}^{[1]},X^{[1]},\mathbf{y}^{[1]},\mathbf{w}^{[1]})\to\mathcal{Q}(\Omega,\Xi,\boldsymbol{\omega}^{[1]},\boldsymbol{\xi}^{[1]},P_1,Q_1,X^{[1]},\mathbf{y}^{[1]},\mathbf{w}^{[1]}),\\ &\mathcal{F}(X,B,X^{[1]},\mathbf{y}^{[1]},\mathbf{w}^{[1]})\to\mathcal{F}(\hat X,\hat B,V,T,P_1,Q_1,X^{[1]},\mathbf{y}^{[1]},\mathbf{w}^{[1]}),\\ &f(X,X^{[1]})\to f(P_1,V,\hat X,X^{[1]}),\quad g(B,\mathbf{y}^{[1]},\mathbf{w}^{[1]})\to g(Q_1,T,\hat B,\mathbf{y}^{[1]},\mathbf{w}^{[1]}).\end{aligned} \tag{3.27}$$

We recall here that $K(X)$ does not depend on P_1, as well, $L(B)$ does not depend on Q_1. Moreover, according to the change (3.26), we have

$$\begin{aligned}\mathcal{P}(\Omega,\Xi,\hat X,\hat B,V,T)=\exp\Big\{-\sum_{j,k}\tilde{\mathfrak{s}}_{jk}Tr\Omega_j\Xi_k\Big\}\cdot\prod_j\frac{1}{\det^M\big(1+M^{-1}V_j^*\hat X_j^{-1}V_j\Omega_jT_j^{-1}\hat B_j^{-1}T_j\Xi_j\big)}\\ \times\prod_{k=p,q}\frac{\det\big(V_k^*\hat X_kV_k+M^{-1}\Omega_kT_k^{-1}\hat B_k^{-1}T_k\Xi_k\big)}{\det\hat B_k}\end{aligned} \tag{3.28}$$

and

$$\begin{aligned}&\mathcal{Q}(\Omega,\Xi,\boldsymbol{\omega}^{[1]},\boldsymbol{\xi}^{[1]},P_1,Q_1,X^{[1]},\mathbf{y}^{[1]},\mathbf{w}^{[1]})\\ &=\prod_{k=p,q}\exp\Big\{-\frac{1}{\sqrt{M}}\sum_j\tilde{\mathfrak{s}}_{jk}\Big(TrP_1^*\Omega_jQ_1\mathbf{w}_k^{[1]}\boldsymbol{\xi}_k^{[1]}J+Tr\boldsymbol{\omega}_k^{[1]}(\mathbf{w}_k^{[1]})^*JQ_1^{-1}\Xi_jP_1\Big)\Big\}\\ &\times\prod_{k,\ell=p,q}\exp\Big\{-\frac{1}{M}\tilde{\mathfrak{s}}_{k\ell}Tr\boldsymbol{\omega}_k^{[1]}(\mathbf{w}_k^{[1]})^*J(\mathbf{w}_\ell^{[1]})\boldsymbol{\xi}_\ell^{[1]}J\Big\}\cdot\prod_{k=p,q}\Big(1-(y_k^{[1]})^{-1}\boldsymbol{\xi}_k^{[1]}(X_k^{[1]})^{-1}\boldsymbol{\omega}_k^{[1]}\Big)^2.\end{aligned} \tag{3.29}$$

Consequently, using (3.25), from (3.17) we can write

$$\begin{aligned}\mathbb{E}|G_{pq,11}(z)|^{2n}=\frac{M^{4W}}{(n!)^28^W\pi^{2W+4}}\int\prod_{j=2}^{W}\mathrm{d}\mu(V_j)\mathrm{d}\nu(T_j)\int_{\mathbb{R}_+^{2W}}\prod_{j=1}^{W}\mathrm{d}b_{j,1}\mathrm{d}b_{j,2}\oint_{\Sigma^{2W}}\prod_{j=1}^{W}\mathrm{d}x_{j,1}\mathrm{d}x_{j,2}\\ \times\exp\Big\{-M\big(K(\hat X,V)+L(\hat B,T)\big)\Big\}\cdot\prod_{j=1}^{W}(x_{j,1}-x_{j,2})^2(b_{j,1}+b_{j,2})^2\cdot\mathsf{A}(\hat X,\hat B,V,T).\end{aligned} \tag{3.30}$$

where we introduced the notation

$$\mathsf{A}(\hat{X},\hat{B},V,T):=\int \mathrm{d}X^{[1]}\mathrm{d}\mathbf{y}^{[1]}\mathrm{d}\mathbf{w}^{[1]}\mathrm{d}\boldsymbol{\omega}^{[1]}\mathrm{d}\boldsymbol{\xi}^{[1]}\mathrm{d}\Omega\mathrm{d}\Xi\mathrm{d}\mu(P_1)\mathrm{d}\nu(Q_1)\ \mathcal{P}(\Omega,\Xi,\hat{X},\hat{B},V,T)$$
$$\times\ \mathcal{Q}(\Omega,\Xi,\boldsymbol{\omega}^{[1]},\boldsymbol{\xi}^{[1]},P_1,Q_1,X^{[1]},\mathbf{y}^{[1]},\mathbf{w}^{[1]})\cdot\mathcal{F}(\hat{X},\hat{B},V,T,P_1,Q_1,X^{[1]},\mathbf{y}^{[1]},\mathbf{w}^{[1]}).\quad(3.31)$$

In (3.30), the regions of V_j's are all $\mathring{U}(2)$, and those of T_j's are all $\mathring{U}(1,1)$. Observe that all Grassmann variables are inside the integrand of the integral $\mathsf{A}(\hat{X},\hat{B},V,T)$. Hence, (3.30) separates the saddle point calculation from the observable $\mathsf{A}(\hat{X},\hat{B},V,T)$.

To facilitate the discussions in the remaining part, we introduce some additional terms and notation here. Henceforth, we will employ the notation

$$(X^{[1]})^{-1}=\Big\{(X_p^{[1]})^{-1},(X_q^{[1]})^{-1}\Big\},\qquad(\mathbf{y}^{[1]})^{-1}=\Big\{(y_p^{[1]})^{-1},(y_q^{[1]})^{-1}\Big\}$$

for the collection of inverse matrices and reciprocals, respectively. For a matrix or a vector A under discussion, we will use the term *A-variables* to refer to all the variables parametrizing it. For example, $\hat{X}_j$-variables means $x_{j,1}$ and $x_{j,2}$, and $\hat{X}$-variables refer to the collection of all $\hat{X}_j$-variables. Analogously, we can define the terms T-variables, $\mathbf{y}^{[1]}$-variables , Ω-variables and so on. We use another term *A-entries* to refer to the non-zero entries of A. Note that $\hat{X}_j$-variables are just $\hat{X}_j$-entries. However, for T_j, they are different, namely,

$$T_j\text{-variables}:\quad t_j,\sigma_j,\quad\text{vs.}\quad T_j\text{-entries}:\quad s_j,t_je^{\mathbf{i}\sigma_j},t_je^{-\mathbf{i}\sigma_j}.$$

Analogously, we will also use the term T-entries to refer to the collection of all T_j-entries. Then V-entries, $\mathbf{w}^{[1]}$-entries, etc. are defined in the same manner. It is easy to check that Q_1^{-1}-entries are the same as Q_1-entries, up to a sign, as well, T_j^{-1}-entries are the same as T_j-entries, for all $j=2,\ldots,W$.

Moreover, to simplify the notation, we make the convention here that we will frequently use a *dot* to represent all the arguments of a function. That means, for instance, we will write $\mathcal{P}(\Omega,\Xi,\hat{X},\hat{B},V,T)$ as $\mathcal{P}(\cdot)$ if there is no confusion. Analogously, we will also use the abbreviation $\mathcal{Q}(\cdot)$, $\mathcal{F}(\cdot)$, $\mathsf{A}(\cdot)$, and so on.

Let $\mathbf{a}:=\{a_1,\ldots,a_\ell\}$ be a set of variables, we will adopt the notation

$$\mathfrak{Q}(\mathbf{a};\kappa_1,\kappa_2,\kappa_3)$$

to denote the class of all multivariate polynomials $\mathfrak{p}(\mathbf{a})$ in the arguments $a_1,\ldots,a_\ell$ such that the following three conditions are satisfied: (i) The total number of the monomials in $\mathfrak{p}(\mathbf{a})$ is bounded by κ_1; (ii) the coefficients of all monomials in $\mathfrak{p}(\mathbf{a})$ are bounded by κ_2 in magnitude; (iii) the power of each a_i in each monomial is bounded by κ_3, for all $i=1,\ldots,\ell$. For example,

$$5b_{j,1}^{-1}+3b_{j,1}t_j^2+1\in\mathfrak{Q}\big(\{b_{j,1}^{-1},b_{j,1},t_j\};3,5,2\big).\quad(3.32)$$

In addition, we define the subset of $\mathfrak{Q}(\mathbf{a};\kappa_1,\kappa_2,\kappa_3)$, namely,

$$\mathfrak{Q}_{\deg}\big(\mathbf{a};\kappa_1,\kappa_2,\kappa_3\big)$$

consisting of those polynomials in $\mathfrak{Q}(\mathbf{a};\kappa_1,\kappa_2,\kappa_3)$ such that the degree is bounded by κ_3, i.e. the total degree of each monomial is bounded by κ_3. For example

$$5b_{j,1}^{-1}+3b_{j,1}t_j^2+1\in\mathfrak{Q}_{\deg}\big(\{b_{j,1}^{-1},b_{j,1},t_j\};3,5,3\big).$$

4. Preliminary discussion on the integrand

In this section, we perform a preliminary analysis on the factors of the integrand in (3.17). For convenience, we introduce the matrix

$$\mathfrak{I} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{4.1}$$

4.1. $\exp\{-M(K(\hat{X},V)+L(\hat{B},T))\}$. Recall the parametrization of $\hat{B}_j$, $\hat{X}_j$, T_j and V_j in (3.22) and (3.24), as well as the matrices defined in (1.24). According to the discussion in [19], there are three types of saddle points of this function, namely,

- Type I : For each j, $(\hat{B}_j, T_j, \hat{X}_j) = (D_\pm, I, D_\pm)$ or $(D_\pm, I, D_\mp)$, $\theta_j \in \mathbb{L}$, $v_j = 0$ if $\hat{X}_j = \hat{X}_1$, and $v_j = 1$ if $\hat{X}_j \neq \hat{X}_1$.
- Type II : For each j, $(\hat{B}_j, T_j, \hat{X}_j) = (D_\pm, I, D_+)$ and $V_j \in \mathring{U}(2)$.
- Type III : For each j, $(\hat{B}_j, T_j, \hat{X}_j) = (D_\pm, I, D_-)$ and $V_j \in \mathring{U}(2)$.

(Actually, since θ_j and v_j vary on continuous sets, it would be more appropriate to use the term *saddle manifolds*.) Note that at each type of saddle points, we have $(\hat{B}_j, T_j) = (D_\pm, I)$ for all j. We will see that the main contribution to the integral (3.17) comes from some small vicinities of the Type I saddle points. Furthermore, the contributions from all the Type I saddle points are the same, which can be explained as follows. At first, by the definition in (3.23), we have $V_1 = I$. If we regard θ_j's in the parametrization of V_j's as fixed parameters, it is easy to see that there are totally 2^W choices of Type I saddle points. Moreover, if $v_j = 1$, we can do the transform

$$\hat{X}_j \to \mathfrak{I}\hat{X}_j\mathfrak{I} = \hat{X}_1, \qquad V_j \to \mathfrak{I}V_j = I \quad \text{in} \quad \mathring{U}(1,1).$$

Consequently, it suffices to consider two saddle points

$$(\hat{B}_j, T_j, \hat{X}_j, V_j) = (D_\pm, I, D_\pm, I), \quad \text{or} \quad (D_\pm, I, D_\mp, I), \tag{4.2}$$

corresponding to $\hat{X}_1 = D_\pm$ or $D_\mp$, respectively. Furthermore, the contributions to the integral (3.17) from the vicinities of these two saddle points are also the same. To see this, we recall the fact that the original integrand in (3.17) is a function of the entries of $X_j = P_j^{-1}\hat{X}_jP_j$. Now we do the transform $P_j \to \mathfrak{I}P_j$ and $\hat{X}_j \to \mathfrak{I}\hat{X}_j\mathfrak{I}$ for all $j = 1, \ldots, W$ to change one saddle in (4.2) to the other. Now, since the Haar measure on $\mathring{U}(2)$ is invariant under the shift $P_1 \to \mathfrak{I}P_1$, the integral over P_1-variables is unchanged. That means, for Type I saddle points, it suffices to consider

- Type I' : For each j, $(\hat{B}_j, T_j, \hat{X}_j, V_j) = (D_\pm, I, D_\pm, I)$.

In summary, the total contribution to the integral (3.17) from all Type I saddle points is 2^W times that from the Type I' saddle point.

Following the discussion in [19], we will show in Section 5 that both $K(\hat{X},V)-K(D_\pm,I)$ and $L(\hat{B},T)-L(D_\pm,I)$ have positive real parts, bounded by some positive quadratic forms from below, which allows us to perform the saddle point analysis. In addition, it will be seen that in a vicinity of Type I' saddle point, $\exp\{-M(K(\hat{X},V)+L(\hat{B},T))\}$ is approximately Gaussian.

4.2. $\mathcal{Q}(\Omega, \Xi, \boldsymbol{\omega}^{[1]}, \boldsymbol{\xi}^{[1]}, P_1, Q_1, X^{[1]}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]})$. The function $\mathcal{Q}(\cdot)$ contains both the Ω, Ξ-variables from $\mathcal{P}(\cdot)$, and the P_1, Q_1, $X^{[1]}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}$-variables from $\mathcal{F}(\cdot)$. In addition, note that in the integrand in (3.17), $\mathcal{Q}(\cdot)$ is the only factor containing the $\boldsymbol{\omega}^{[1]}$ and $\boldsymbol{\xi}^{[1]}$- variables. Hence, we can compute the integral

$$\mathsf{Q}\big(\Omega, \Xi, P_1, Q_1, X^{[1]}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}\big) := \int d\boldsymbol{\omega}^{[1]} d\boldsymbol{\xi}^{[1]}\ \mathcal{Q}\big(\Omega, \Xi, \boldsymbol{\omega}^{[1]}, \boldsymbol{\xi}^{[1]}, P_1, Q_1, X^{[1]}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}\big) \tag{4.3}$$

at first. The explicit formula for $\mathsf{Q}(\cdot)$ is complicated and irrelevant for us. From (3.29) and the definition of the Grassmann integral, it is not difficult to see that $\mathsf{Q}(\cdot)$ is a polynomial of the $(X^{[1]})^{-1}$, $(\mathbf{y}^{[1]})^{-1}$,

$\mathbf{w}^{[1]}$, P_1, Q_1, Ω and Ξ-entries. In principle, for each monomial in the polynomial $\mathsf{Q}(\cdot)$, we can combine the Grassmann variables with $\mathcal{P}(\cdot)$, then perform the integral over Ω and Ξ-variables, whilst we combine the complex variables with $\mathcal{F}(\cdot)$, and perform the integral over $X^{[1]}$, $\mathbf{y}^{[1]}$, $\mathbf{w}^{[1]}$, P_1 and Q_1-variables. A formal discussion on $\mathsf{Q}(\cdot)$ will be given in Section 6.1. However, the terms from $\mathsf{Q}(\cdot)$ turn out to be irrelevant in our proof. Therefore, in the arguments with $\mathsf{Q}(\cdot)$ involved, a typical strategy that we will adopt is as follows: we usually neglect $\mathsf{Q}(\cdot)$ at first, and perform the discussion on $\mathcal{P}(\cdot)$ and $\mathcal{F}(\cdot)$ separately, at the end, we make necessary comments on how to slightly modify the discussions to take $\mathsf{Q}(\cdot)$ into account.

4.3. $\mathcal{P}(\Omega, \Xi, \hat{X}, \hat{B}, V, T)$**.** We will mainly regard $\mathcal{P}(\cdot)$ as a function of the Ω and Ξ-variables. As mentioned above, we also have some Ω and Ξ-variables from the irrelevant term $\mathsf{Q}(\cdot)$. But we temporarily ignore them and regard as if the integral over Ω and Ξ-variables reads

$$\mathsf{P}(\hat{X}, \hat{B}, V, T) := \int d\Omega d\Xi \, \mathcal{P}(\Omega, \Xi, \hat{X}, \hat{B}, V, T). \tag{4.4}$$

We shall estimate $\mathsf{P}(\cdot)$ in three different regions: (1) the complement of the vicinities of the saddle points; (2) the vicinity of Type I saddle point; (3) the vicinities of Type II and III saddle points, which will be done in Sections 6.2, 9.1 and 10.1, respectively. (Definition 5.5 gives the precise definition of the vicinities.) In each case we will decompose the function $\mathcal{P}(\cdot)$ as a product of a Gaussian measure and a multivariate polynomial of Grassmann variables. Consequently, we can employ (3.2) to perform the integral of this polynomial against the Gaussian measure, whereby $\mathsf{P}(\cdot)$ can be estimated. Especially, it turns out that in the vicinity of Type I saddle points, $\mathsf{P}(\cdot)$ is approximately the normalizing constant of the Gaussian measure obtained from $\exp\{-M(K(\hat{X}, V) + L(\hat{B}, T))\}$.

4.4. $\mathcal{F}(\hat{X}, \hat{B}, V, T, P_1, Q_1, X^{[1]}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]})$**.** Observe that $\mathcal{F}$ is the only term containing the energy scale η. As in the previous discussion of $\mathcal{P}(\cdot)$, here we also ignore the P_1, Q_1, $X^{[1]}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}$-variables from the irrelevant term $\mathsf{Q}(\cdot)$ temporarily, and investigate the integral

$$\begin{aligned}\mathsf{F}(\hat{X}, \hat{B}, V, T) &= \int dX^{[1]} d\mathbf{y}^{[1]} \mathrm{d}\mathbf{w}^{[1]} d\mu(P_1) d\nu(Q_1) \, \mathcal{F}(\hat{X}, \hat{B}, V, T, P_1, Q_1, X^{[1]}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}) \\ &= \int dX^{[1]} d\mu(P_1) \, f(\hat{X}, V, P_1) \cdot \int d\mathbf{y}^{[1]} \mathrm{d}\mathbf{w}^{[1]} d\nu(Q_1) \, g(\hat{B}, T, Q_1, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}). \end{aligned} \tag{4.5}$$

We shall also estimate $\mathsf{F}(\cdot)$ in three different regions: (1) the complement of the vicinities of the saddle points; (2) the vicinity of Type I saddle point; (3) the vicinities of Type II and III saddle points, which will be done in Sections 6.3, 9.2 and 10.2, respectively.

Especially, when we restrict the $\hat{X}$, $\hat{B}$, V and T-variables to the vicinity of the Type I saddle points, the above integral can be performed approximately, resulting our main term, a factor of order $1/(N\eta)^{n+2}$. This step will be done in Section 9. It is instructive to give a heuristic sketch of this calculation. At first, we plug the Type I saddle points into (4.5). We will show that the integral of $f(\cdot)$ approximately reads

$$e^{-(a_+ - a_-)N\eta} \int dX^{[1]} d\mu(P_1) \, f(D_\pm, I, P_1) \sim \frac{1}{N\eta},$$

which is the easy part. Then, recalling the definition of $g(\cdot)$ in (3.20) and the parameterization (3.15), we will show that the integral of $g(\cdot)$ approximately reads

$$\begin{aligned} & e^{(a_+ - a_-)N\eta} \int d\mathbf{y}^{[1]} \mathrm{d}\mathbf{w}^{[1]} d\nu(Q_1) \, g(D_\pm, I, Q_1, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}) \\ & \sim \int_0^\infty 2t dt \int_{\mathbb{L}^2} d\sigma_p^{[1]} d\sigma_q^{[1]} \, e^{\mathbf{i}n\sigma_p^{[1]}} e^{-\mathbf{i}n\sigma_q^{[1]}} \cdot e^{-cN\eta t^2 + c_1 e^{-\mathbf{i}\sigma_p^{[1]}} t + c_2 e^{\mathbf{i}\sigma_q^{[1]}} t} \\ & \sim \int_0^\infty 2t dt \cdot t^{2n} \cdot e^{-cN\eta t^2} \sim \frac{1}{(N\eta)^{n+1}}, \end{aligned} \tag{4.6}$$

where in the second step above we used the fact that

$$\int_{\mathbb{L}} d\sigma \cdot e^{\mathbf{i}n\sigma}\, e^{ce^{-\mathbf{i}\sigma}t} \sim t^n.$$

We notice that the factor $e^{\mathbf{i}n\sigma_p^{[1]}}e^{-\mathbf{i}n\sigma_q^{[1]}}$ in (4.6) actually comes from the term

$$\Big(\big(\mathbf{w}_q^{[1]}(\mathbf{w}_q^{[1]})^*\big)_{12}\big(\mathbf{w}_p^{[1]}(\mathbf{w}_p^{[1]})^*\big)_{21}\Big)^n$$

in (3.20). This factor brings a strong oscillation to the integrand in the integral (4.6). In Case 2, an analogous factor will appear, resulting the same estimate as (4.6). However, in Case 3, such an oscillating factor is absent, then the estimate for the counterpart of the integral in (4.6) is of order $1/N\eta$ instead of $1/(N\eta)^{n+1}$. The detailed analysis will be presented in Sections 10 and 11.

5. Saddle points and vicinities

In this section, we study the saddle points of $K(\hat{X}, V)$ and $L(\hat{B}, T)$ and deform the contours of the $\hat{B}$-variables to pass through the saddle points. Then we introduce and classify some small vicinities of these saddle points. The derivation of the saddle points of $K(\hat{X}, V)$ and $L(\hat{B}, T)$ in Section 5.1 and 5.2 below is essentially the same as the counterpart in [19], the only difference is that we are working under a more general setting on S. Hence, in Section 5.1 and 5.2, we just sketch the discussion, list the results, and make necessary modifications to adapt to our setting. In the sequel, we employ the notation

$$\begin{aligned}&\mathbf{b}_a := (b_{1,a}, \dots, b_{W,a}), \quad \mathbf{x}_a := (x_{1,a}, \dots, x_{W,a}), \qquad a = 1, 2,\\ &\mathbf{t} := (t_2, \dots, t_W), \quad \mathbf{v} := (v_2, \dots, v_W), \quad \boldsymbol{\sigma} := (\sigma_2, \dots, \sigma_W), \quad \boldsymbol{\theta} := (\theta_2, \dots, \theta_W).\end{aligned} \tag{5.1}$$

As mentioned above, later we also need to deform the contours, and discuss the integral over some vicinities of the saddle points, thus it is convenient to introduce a notation for the integral over specific domains. To this end, for $a = 1, 2$, we use $\mathbf{I}_a^b$ and $\mathbf{I}_a^x$ to denote generic domains of $\mathbf{b}_a$ and $\mathbf{x}_a$ respectively. Analogously, we use $\mathbf{I}^t$ and $\mathbf{I}^v$ to represent generic domains of $\mathbf{t}$ and $\mathbf{v}$, respectively. These domains will be specified later. Now, for some collection of domains, we introduce the notation

$$\begin{aligned}&\mathcal{I}(\mathbf{I}_1^b, \mathbf{I}_2^b, \mathbf{I}_1^x, \mathbf{I}_2^x, \mathbf{I}^t, \mathbf{I}^v) := \frac{M^{4W}}{(n!)^2 8^W \pi^{2W+4}} \int_{\mathbb{L}^{2W-2}} \prod_{j=2}^W \frac{\mathrm{d}\theta_j}{2\pi} \prod_{j=2}^W \frac{\mathrm{d}\sigma_j}{2\pi}\\ &\times \int_{\mathbf{I}_1^b} \prod_{j=1}^W \mathrm{d}b_{j,1} \int_{\mathbf{I}_2^b} \prod_{j=1}^W \mathrm{d}b_{j,2} \int_{\mathbf{I}_1^x} \prod_{j=1}^W \mathrm{d}x_{j,1} \int_{\mathbf{I}_2^x} \prod_{j=1}^W \mathrm{d}x_{j,2} \int_{\mathbf{I}^t} \prod_{j=2}^W 2t_j \mathrm{d}t_j \int_{\mathbf{I}^v} \prod_{j=2}^W 2v_j \mathrm{d}v_j\\ &\times \exp\Big\{-M\big(K(\hat{X}, V) + L(\hat{B}, T)\big)\Big\} \cdot \prod_{j=1}^W (x_{j,1} - x_{j,2})^2 (b_{j,1} + b_{j,2})^2 \cdot \mathsf{A}(\hat{X}, \hat{B}, V, T).\end{aligned} \tag{5.2}$$

For example, we can write (3.30) as

$$\mathbb{E}|G_{pq,11}(z)|^{2n} = \mathcal{I}(\mathbb{R}_+^W, \mathbb{R}_+^W, \Sigma^W, \Sigma^W, \mathbb{R}_+^{W-1}, \mathbb{I}^{W-1}), \tag{5.3}$$

which is the integral over the full domain.

5.1. Saddle points of $L(\hat{B}, T)$.

We introduce the function

$$\Bbbk(a) := \frac{a^2}{2} - \mathbf{i}Ea - \log a, \quad a \in \mathbb{C}. \tag{5.4}$$

Recalling the definition of $L(\cdot)$ in (3.18), the decomposition of B_j's in (3.21) and the definition of T_j's in (3.23), we can write

$$L(\hat{B}, T) = -\frac{1}{4}\sum_{j,k} \mathfrak{s}_{jk} Tr\big(T_j^{-1}\hat{B}_j T_j - T_k^{-1}\hat{B}_k T_k\big)^2 + \sum_j \Big(\frac{1}{2} Tr\hat{B}_j^2 - \mathbf{i}ETr\hat{B}_j - \log\det\hat{B}_j\Big)$$

$$= -\frac{1}{4}\sum_{j,k}\mathfrak{s}_{jk}Tr(\hat{B}_j-\hat{B}_k)^2+\sum_j\Big(\frac{1}{2}Tr\hat{B}_j^2-\mathbf{i}ETr\hat{B}_j-\log\det\hat{B}_j\Big)$$
$$+\frac{1}{2}\sum_{j,k}\mathfrak{s}_{jk}|(T_kT_j^{-1})_{12}|^2(b_{j,1}+b_{j,2})(b_{k,1}+b_{k,2})$$
$$=:\ell(\mathbf{b}_1)+\ell(-\mathbf{b}_2)+\ell_S(\hat{B},T), \tag{5.5}$$

where we used the notation introduced in (5.1), and the functions $\ell(\cdot)$ and $\ell_S(\cdot)$ are defined as

$$\ell(\mathbf{a}):=-\frac{1}{4}\sum_{j,k}\mathfrak{s}_{jk}(a_j-a_k)^2+\sum_j\Bbbk(a_j),\quad \mathbf{a}=(a_1,\ldots,a_W)\in\mathbb{C}^W,$$
$$\ell_S(\hat{B},T):=\frac{1}{2}\sum_{j,k}\mathfrak{s}_{jk}|(T_kT_j^{-1})_{12}|^2(b_{j,1}+b_{j,2})(b_{k,1}+b_{k,2}). \tag{5.6}$$

Following the discussion in [19] with slight modification (see Section 3 therein), we see that for $|E|\le\sqrt{2}-\kappa$, the saddle point of $L(\hat{B},T)$ is

$$(\hat{B}_j,T_j)=(D_\pm,I),\quad \forall\, j=1,\ldots,W, \tag{5.7}$$

where $D_\pm$ is defined in (1.24). For simplicity, we will write (5.7) as $(\hat{B},T)=(D_\pm,I)$ in the sequel. Observe that

$$L(D_\pm,I)=\ell(a_+)+\ell(a_-),\qquad \ell_S(D_\pm,I)=0. \tag{5.8}$$

We introduce the notation

$$\mathring{\ell}_{++}(\mathbf{a}):=\ell(\mathbf{a})-\ell(a_+),\quad \mathring{\ell}_{+-}(\mathbf{a}):=\ell(\mathbf{a})-\ell(a_-)\quad \mathring{\ell}_{--}(\mathbf{a}):=\ell(-\mathbf{a})-\ell(a_-), \tag{5.9}$$

where $\ell(a_+)$ represents the value of $\ell(\mathbf{a})$ at the point $\mathbf{a}=(a_+,\ldots,a_+)$, and $\ell(a_-)$ is defined analogously. Correspondingly, we adopt the notation

$$\mathring{L}(\hat{B},T):=L(\hat{B},T)-L(D_\pm,I)=\mathring{\ell}_{++}(\mathbf{b}_1)+\mathring{\ell}_{--}(\mathbf{b}_2)+\ell_S(\hat{B},T), \tag{5.10}$$

which is implied by (5.5), (5.8) and (5.9). Now, for each $j=1,\ldots,W$, we deform the contours of $b_{j,1}$ and $b_{j,2}$ to

$$b_{j,1}\in\Gamma:=\{ra_+|r\in\mathbb{R}_+\},\quad b_{j,2}\in\bar{\Gamma}=\{-ra_-|r\in\mathbb{R}_+\} \tag{5.11}$$

to pass through the saddle points of $\hat{B}$-variables, based on the following lemma which will be proved in Section 7.

Lemma 5.1. *With the notation introduced in (5.2), we have*

$$\mathcal{I}\big(\Gamma^W,\bar{\Gamma}^W,\Sigma^W,\Sigma^W,\mathbb{R}_+^{W-1},\mathbb{I}^{W-1}\big)=\mathcal{I}\big(\mathbb{R}_+^W,\mathbb{R}_+^W,\Sigma^W,\Sigma^W,\mathbb{R}_+^{W-1},\mathbb{I}^{W-1}\big)=\mathbb{E}|G_{pq,11}(z)|^{2n}.$$

We introduce the notation

$$r_{j,1}=|b_{j,1}|,\quad r_{j,2}=|b_{j,2}|,\quad j=1,\ldots,W. \tag{5.12}$$

Along the new contours, we have the following lemma.

Lemma 5.2. *Suppose that $|E|\le\sqrt{2}-\kappa$. Let $\mathbf{b}_1\in\Gamma^W$, $\mathbf{b}_2\in\bar{\Gamma}^W$. We have*

$$\mathsf{Re}\mathring{L}(\hat{B},T)\ge c\sum_{a=1,2}\sum_{j=1}^W\Big((r_{j,a}-1)^2+(r_{j,a}-\log r_{j,a}-1)\Big)+\mathsf{Re}\ell_S(\hat{B},T)\ge c\sum_{a=1,2}\sum_{j=1}^W(r_{j,a}-1)^2 \tag{5.13}$$

for some positive constant c.

Proof. Since $|E| \le \sqrt{2} - \kappa$, we have $\mathsf{Re}(b_{j,1} + b_{j,2})(b_{k,1} + b_{k,2}) \ge 0$ for $\mathbf{b}_1 \in \Gamma^W$ and $\mathbf{b}_2 \in \bar{\Gamma}^W$, thus $\mathsf{Re}\ell_S(\hat{B}, T) \ge 0$, in light of the definition in (5.6). Consequently, according to (5.10), it suffices to prove

$$\mathsf{Re}\mathring{\ell}_{++}(\mathbf{b}_1) + \mathsf{Re}\mathring{\ell}_{--}(\mathbf{b}_2) \ge c \sum_{a=1,2} \sum_{j=1}^{W} \Big((r_{j,a} - 1)^2 + (r_{j,a} - \log r_{j,a} - 1) \Big) \tag{5.14}$$

for some positive constant c. To see this, we observe the following identities obtained via elementary calculation,

$$\mathsf{Re}\mathring{\ell}_{++}(\mathbf{b}_1) = \frac{E^2 - 2}{4} \Big(\frac{1}{2} \sum_{j,k} \mathfrak{s}_{jk} (r_{j,1} - r_{k,1})^2 - \sum_j (r_{j,1} - 1)^2 \Big) + \sum_{j=1}^{W} \big(r_{j,1} - \log r_{j,1} - 1 \big)$$

$$\mathsf{Re}\mathring{\ell}_{--}(\mathbf{b}_2) = \frac{E^2 - 2}{4} \Big(\frac{1}{2} \sum_{j,k} \mathfrak{s}_{jk} (r_{j,2} - r_{k,2})^2 - \sum_j (r_{j,2} - 1)^2 \Big) + \sum_{j=1}^{W} \big(r_{j,2} - \log r_{j,2} - 1 \big),$$

which together with $|E| \le \sqrt{2} - \kappa$ and (1.4) implies (5.14) immediately. Hence, we completed the proof of Lemma 5.2. □

5.2. **Saddle points of $K(\hat{X}, V)$.** Analogously, recalling the definition in (5.6), we can write

$$\begin{aligned} K(\hat{X}, V) &= \frac{1}{4} \sum_{j,k} \mathfrak{s}_{jk} Tr \big(V_j^* \hat{X}_j V_j - V_k^* \hat{X}_k V_k \big)^2 - \sum_j \Big(\frac{1}{2} Tr \hat{X}_j^2 - \mathbf{i} E Tr \hat{X}_j - \log\det \hat{X}_j \Big) \\ &= \frac{1}{4} \sum_{j,k} \mathfrak{s}_{jk} Tr \big(\hat{X}_j - \hat{X}_k \big)^2 - \sum_j \Big(\frac{1}{2} Tr \hat{X}_j^2 - \mathbf{i} E Tr \hat{X}_j - \log\det \hat{X}_j \Big) \\ &\quad + \frac{1}{2} \sum_{j,k} \mathfrak{s}_{jk} |(V_k V_j^*)_{12}|^2 (x_{j,1} - x_{j,2})(x_{k,1} - x_{k,2}) \\ &= -\ell(\mathbf{x}_1) - \ell(\mathbf{x}_2) + \ell_S(\hat{X}, V), \end{aligned} \tag{5.15}$$

where $\ell(\cdot)$ is defined in the first line of (5.6) and $\ell_S(\hat{X}, V)$ is defined as

$$\ell_S(\hat{X}, V) = \frac{1}{2} \sum_{j,k} \mathfrak{s}_{jk} |(V_k V_j^*)_{12}|^2 (x_{j,1} - x_{j,2})(x_{k,1} - x_{k,2}). \tag{5.16}$$

Analogously to the notation $L(D_\pm, I)$, we will use $K(D_\pm, I)$ to represent the value of $K(\hat{X}, V)$ at $(\hat{X}_j, V_j) = (D_\pm, I)$ for all $j = 1, \dots, W$. In addition, $K(D_+, I)$ and $K(D_-, I)$ are defined in the same manner. Observing that

$$\ell_S(D_\pm, I) = \ell_S(D_+, I) = \ell_S(D_-, I) = 0, \tag{5.17}$$

we have

$$K(D_\pm, I) = -\ell(a_+) - \ell(a_-), \quad K(D_+, I) = -2\ell(a_+), \quad K(D_-, I) = -2\ell(a_-). \tag{5.18}$$

Moreover, we employ the notation

$$\mathring{K}(\hat{X}, V) = K(\hat{X}, V) - K(D_\pm, I) = -\mathring{\ell}_{++}(\mathbf{x}_1) - \mathring{\ell}_{+-}(\mathbf{x}_2) + \ell_S(\hat{X}, V). \tag{5.19}$$

We will need the following elementary observations that are easy to check from (5.18) and (5.6)

$$K(D_\pm, I) + L(D_\pm, I) = 0, \qquad \mathsf{Re}K(D_+, I) = \mathsf{Re}K(D_-, I) = \mathsf{Re}K(D_\pm, I). \tag{5.20}$$

In addition, we introduce the $W \times W$ matrix

$$S^v = (\mathfrak{s}^v_{jk}), \quad \mathfrak{s}^v_{jk} := \mathfrak{s}_{jk} |(V_k V_j^*)_{12}|^2, \tag{5.21}$$

and the $2W \times 2W$ matrices

$$\mathbb{S} = S \oplus S, \qquad \mathbb{S}^v := \mathbb{S} + \begin{pmatrix} -S^v & S^v \\ S^v & -S^v \end{pmatrix}, \tag{5.22}$$

where $\mathbb{S}^v$ depends on the V-variables according to (5.21). Here we regard V-variables as fixed parameters. Due to the fact $|(V_k V_j^*)_{12}| \in \mathbb{I}$, it is easy to see that $\mathbb{S}^v$ is a weighted Laplacian of a graph with $2W$ vertices. In particular, $\mathbb{S}^v \le 0$. By the definition (5.21), one can see that $S^v_{ii} = 0$ for all $i = 1, \dots, W$. Consequently, we can obtain

$$\sum_{k \neq j} \mathbb{S}^v_{jk} = \sum_{k \neq j} \mathbb{S}^v_{kj} = -\mathbb{S}^v_{jj} = \begin{cases} -\mathfrak{s}_{jj}, & \text{if} \quad j = 1, \dots W \\ -\mathfrak{s}_{j-W, j-W}, & \text{if} \quad j = W+1, \dots, 2W \end{cases}$$

Similarly to (1.4), we get

$$I + \mathbb{S}^v \ge c_0 I, \tag{5.23}$$

where c_0 is the constant in Assumption 1.1 (ii). Moreover, it is not difficult to see from the definitions in (5.16), (5.21) and (5.22) that

$$\frac{1}{4} \sum_{j,k} \mathfrak{s}_{jk} Tr(\hat{X}_j - \hat{X}_k)^2 + \ell_S(\hat{X}, V) = -\frac{1}{2} \mathbf{x}' \mathbb{S}^v \mathbf{x}, \tag{5.24}$$

where we used the notation $\mathbf{x} := (\mathbf{x}_1', \mathbf{x}_2')'$.

Now let

$$\vartheta_j = \arg x_{j,1}, \quad \vartheta_{W+j} = \arg x_{j,2}, \quad \forall\, j = 1, \dots, W. \tag{5.25}$$

Then, recalling the parametrization of V_j's in (3.24), we have the following lemma.

Lemma 5.3. *Assume that $x_{j,1}, x_{j,2} \in \Sigma$ for all $j = 1, \dots, W$. We have*

$$\mathring{\mathsf{Re}K}(\hat{X}, V) \ge \frac{1}{4} \sum_{j,k=1}^{2W} (\mathbb{S}^v)_{jk} (\cos \vartheta_j - \cos \vartheta_k)^2 + c \sum_{j=1}^{2W} \Big(\sin \vartheta_j - \frac{E}{2} \Big)^2 \tag{5.26}$$

for some positive constant c. In addition, $\mathring{\mathsf{Re}K}(\hat{X}, V)$ attains its minimum 0 at the following three types of saddle points

- *Type I :* *For each j, $\hat{X}_j = D_\pm$ or $D_\mp$,*
  $\theta_j \in \mathbb{L}$ $v_j = 0$ if $\hat{X}_j = \hat{X}_1$, and $v_j = 1$ if $\hat{X}_j \neq \hat{X}_1$,
- *Type II :* *For each j, $\hat{X}_j = D_+, V_j \in \mathring{U}(2)$,*
- *Type III :* *For each j, $\hat{X}_j = D_-, V_j \in \mathring{U}(2)$,*

which are the restrictions of three types of saddle points in Section 4.1, on $\hat{X}$ and V-variables.

Remark 5.4. *The Type I saddle points of $(\hat{X}, V)$ are exactly those points satisfying*

$$V_j^* \hat{X}_j V_j = D_\pm, \quad \forall\, j = 1, \dots, W, \quad \text{or} \quad V_j^* \hat{X}_j V_j = D_\mp, \quad \forall\, j = 1, \dots, W.$$

In Lemma 5.3, we wrote them in terms of $\hat{X}_j$, v_j and θ_j in order to evoke the parameterization in (3.22) and (3.24).

Proof. By (5.15), (5.24), the definitions of the functions $\ell(\cdot)$ in (5.7) and $\mathbb{k}(\cdot)$ in (5.4), we can write

$$\mathring{K}(\hat{X}, V) = -\frac{1}{2} \mathbf{x}^* \mathbb{S}^v \mathbf{x} - \sum_{j=1}^{W} \Big(\mathbb{k}(x_{j,1}) + \mathbb{k}(x_{j,2}) - \mathbb{k}(a_+) - \mathbb{k}(a_-) \Big).$$

By using (5.25) and the fact $|x_{j,a}| = 1$ for all $j = 1, \dots, W$ and $a = 1, 2$, we can obtain via elementary calculation

$$\begin{aligned}\mathring{\mathsf{Re}K}(\hat{X}, V) = &\frac{1}{4}\sum_{j,k=1}^{2W}(\mathbb{S}^v)_{jk}(\cos\vartheta_j - \cos\vartheta_k)^2 \\ &-\frac{1}{4}\sum_{j,k=1}^{2W}(\mathbb{S}^v)_{jk}(\sin\vartheta_j - \sin\vartheta_k)^2 + \sum_{j=1}^{2W}\Big(\sin\vartheta_j - \frac{E}{2}\Big)^2. \end{aligned} \tag{5.27}$$

In light of the fact $\mathbb{S}^v \le 0$ and (5.23), we have

$$I + \frac{1}{2}\mathbb{S}^v \ge I + \mathbb{S}^v \ge c_0 I. \tag{5.28}$$

Applying (5.28) to the last two terms on the r.h.s. of (5.27) yields (5.26).

Now, we show that $\mathring{\mathsf{Re}K}(\hat{X}, V)$ attains its minimum 0 at three types of points listed in Lemma 5.3. It is elementary to check that these points are minimum points along the contour by plugging them into the definition of $\mathring{\mathsf{Re}K}(\hat{X}, V)$. In the sequel, we show that they are the only solutions to the equation

$$\mathring{\mathsf{Re}K}(\hat{X}, V) = 0. \tag{5.29}$$

At first, by the second term on the r.h.s. of (5.26), we see that for any solution to (5.29),

$$\sin\vartheta_j = \frac{E}{2}, \quad \forall\, j = 1, \dots, 2W,$$

which implies that $x_{j,a} = a_+$ or a_- for all $j = 1, \dots, W$ and $a = 1, 2$, by recalling the definition (5.25) and the definitions of a_+ and a_- in Section 1.4. Consequently, for each j, $\hat{X}_j$ can only be one of $D_\pm$, $D_\mp$, D_+ and D_-.

Suppose that $\hat{X}_1 = D_+$, we claim that $\hat{X}_j = D_+$ for all j. Otherwise, owing to the fact that the graph $\mathcal{G}$ is connected, there exists $\{i, j\} \in \mathcal{E}$ such that $\mathfrak{s}_{ij} > 0$ and $\hat{X}_i = D_+$ but $\hat{X}_j = D_\pm$, $D_\mp$ or D_-. Without loss of generality, we assume $\hat{X}_j = D_\pm$. In this case, we use the fact

$$\begin{aligned}\mathring{\mathsf{Re}K}(\hat{X}, V) \ge &\frac{1}{4}(\mathbb{S}^v)_{ij}(\cos\vartheta_i - \cos\vartheta_j)^2 + \frac{1}{4}(\mathbb{S}^v)_{i+W,j}(\cos\vartheta_{i+W} - \cos\vartheta_j)^2 \\ &+ \frac{1}{4}(\mathbb{S}^v)_{i,j+W}(\cos\vartheta_i - \cos\vartheta_{j+W})^2 + \frac{1}{4}(\mathbb{S}^v)_{i+W,j+W}(\cos\vartheta_{i+W} - \cos\vartheta_{j+W})^2, \end{aligned} \tag{5.30}$$

which follows from (5.26) directly. Now, by the assumption $\hat{X}_i = D_+$ while $\hat{X}_j = D_\pm$, we have

$$\cos\vartheta_i = \cos\vartheta_{i+W} = \cos\vartheta_j = \mathsf{Re}(a_+) = \frac{\sqrt{4 - E^2}}{2}, \quad \cos\vartheta_{j+W} = \mathsf{Re}(a_-) = -\frac{\sqrt{4 - E^2}}{2},$$

which together with (5.30) implies that

$$\mathring{\mathsf{Re}K}(\hat{X}, V) \ge \frac{4 - E^2}{4}\Big((\mathbb{S}^v)_{i,j+W} + (\mathbb{S}^v)_{i+W,j+W}\Big) = \frac{4 - E^2}{4}\mathfrak{s}_{ij} > 0, \tag{5.31}$$

contradicting to (5.29). Analogously, we can show that $\hat{X}_j$ can not be $D_\mp$ or D_-. Consequently, for a solution to (5.29), if $\hat{X}_1 = D_+$, we have shown that $\hat{X}_j = D_+$ for all j. Similarly, we can show that if $\hat{X}_1 = D_-$, then $\hat{X}_j = D_-$ for all j. These two kinds of solutions are collected as the Type II and Type III saddle points, respectively.

What remains is to show that if $\hat{X}_1 = D_\pm$ or $D_\mp$, the solution to (5.29) must be one of the Type I saddle points. We only show the case of $\hat{X}_1 = D_\pm$. Assume that $\{1, i\} \in \mathcal{E}$ in the graph $\mathcal{G}$, i.e. $\mathfrak{s}_{1i} > 0$. At first, similarly to the discussion from (5.30) to (5.31), we can show that $\hat{X}_i$ can only be $D_\pm$ or $D_\mp$. If $\hat{X}_i = D_\pm$, then by using (5.30) with $j = 1$, we have

$$\mathring{\mathsf{Re}K}(\hat{X}, V) \ge \frac{4 - E^2}{4}\Big((\mathbb{S}^v)_{i,1+W} + (\mathbb{S}^v)_{i+W,1}\Big) = \frac{4 - E^2}{2}\mathfrak{s}^v_{1i} \ge 0,$$

and the equality holds if and only if $V_i = I$, according to the assumption $V_1 = I$ and the definition in (5.21). The discussion on the case of $\hat{X}_i = D_{\mp}$ is analogous. Consequently, we have

$$V_i^* \hat{X}_i V_i = V_1^* \hat{X}_1 V_1 = D_{\pm}. \tag{5.32}$$

Since the graph $\mathcal{G}$ is connected, we can show that (5.32) holds for all $i = 1, \dots, W$. Analogously, if $\hat{X}_1 = D_{\mp}$, we can show that $V_j^* \hat{X}_j V_j = D_{\mp}$ for all $j = 1, \dots, W$. Therefore, we completed the proof of Lemma 5.3. □

5.3. **Vicinities of the saddle points.** Having studied the saddle points of $L(\hat{B}, T)$ and $K(\hat{X}, V)$, we then introduce some small vicinities of them. To this end, we introduce the quantity

$$\Theta \equiv \Theta(N, \varepsilon_0) := W N^{\varepsilon_0} \tag{5.33}$$

for small positive constant ε_0 which will be chosen later. Let $\mathbf{a} = (a_1, \dots, a_W) \in \mathbb{C}^W$ be any complex vector. In the sequel, we adopt the notation

$$\begin{aligned} \mathbf{a} + d := (a_1 + d, \dots, a_W + d), \quad d\mathbf{a} := (da_1, \dots, da_W), \quad \forall\, d \in \mathbb{C}, \\ \arg(\mathbf{a}) := \big(\arg(a_1), \dots, \arg(a_W)\big). \end{aligned} \tag{5.34}$$

Now, we define the following domains .

$$\begin{aligned} \Upsilon_+^b \equiv \Upsilon_+^b(N, \varepsilon_0) &:= \Big\{ \mathbf{a} \in \Gamma^W : ||\mathbf{a} - a_+||_2^2 \le \frac{\Theta}{M} \Big\}, \\ \Upsilon_-^b \equiv \Upsilon_-^b(N, \varepsilon_0) &:= \Big\{ \mathbf{a} \in \bar{\Gamma}^W : ||\mathbf{a} + a_-||_2^2 \le \frac{\Theta}{M} \Big\}, \\ \Upsilon_+^x \equiv \Upsilon_+^x(N, \varepsilon_0) &:= \Big\{ \mathbf{a} \in \Sigma^W : ||\arg(a_+^{-1}\mathbf{a})||_2^2 \le \frac{\Theta}{M} \Big\}, \\ \Upsilon_-^x \equiv \Upsilon_-^x(N, \varepsilon_0) &:= \Big\{ \mathbf{a} \in \Sigma^W : ||\arg(a_-^{-1}\mathbf{a})||_2^2 \le \frac{\Theta}{M} \Big\}, \\ \Upsilon_S \equiv \Upsilon_S(N, \varepsilon_0) &:= \Big\{ \mathbf{a} \in \mathbb{R}_+^{W-1} : -\mathbf{a}' S^{(1)} \mathbf{a} \le \frac{\Theta}{M} \Big\}, \end{aligned} \tag{5.35}$$

where the superscripts b and x indicate that these will be domains of the corresponding variables. In order to define the vicinities of the Type I saddle points properly, we introduce the permutation ϵ_j of $\{1, 2\}$, for each triple $(x_{j,1}, x_{j,2}, v_j)$. Specifically, recalling the fact of $u_j = \sqrt{1 - v_j^2}$ from (3.24), we define

$$v_{j,\epsilon_j} \equiv v_{j,\epsilon_j}(\epsilon_1) := v_j \mathbf{1}(\epsilon_j = \epsilon_1) + u_j \mathbf{1}(\epsilon_j \neq \epsilon_1).$$

Denoting by $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_W)$ and $\boldsymbol{\epsilon}(a) = (\epsilon_1(a), \dots, \epsilon_W(a))$ for $a = 1, 2$, we set

$$\mathbf{x}_{\boldsymbol{\epsilon}(a)} = (x_{1,\epsilon_1(a)}, \dots, x_{W,\epsilon_W(a)}), \qquad a = 1, 2, \qquad \mathbf{v}_{\boldsymbol{\epsilon}} = (v_{2,\epsilon_2}, \dots, v_{W,\epsilon_W}). \tag{5.36}$$

With this notation, we now define the Type I, II, and III vicinities, parameterized by $(\mathbf{b}_1, \mathbf{b}_2, \mathbf{x}_1, \mathbf{x}_2, \mathbf{t}, \mathbf{v})$ of the corresponding saddle point types. We also define the special case of the Type I vicinity, namely, Type I' vicinity, corresponding to the Type I' saddle point defined in Section 4.1.

Definition 5.5.

- *Type I vicinity :* $\big(\mathbf{b}_1, \mathbf{b}_2, \mathbf{x}_{\boldsymbol{\epsilon}(1)}, \mathbf{x}_{\boldsymbol{\epsilon}(2)}, \mathbf{t}, \mathbf{v}_{\boldsymbol{\epsilon}}\big) \in \Upsilon_+^b \times \Upsilon_-^b \times \Upsilon_+^x \times \Upsilon_-^x \times \Upsilon_S \times \Upsilon_S$ *for some* $\boldsymbol{\epsilon}$.
- *Type I' vicinity :* $\big(\mathbf{b}_1, \mathbf{b}_2, \mathbf{x}_1, \mathbf{x}_2, \mathbf{t}, \mathbf{v}\big) \in \Upsilon_+^b \times \Upsilon_-^b \times \Upsilon_+^x \times \Upsilon_-^x \times \Upsilon_S \times \Upsilon_S$.
- *Type II vicinity :* $\big(\mathbf{b}_1, \mathbf{b}_2, \mathbf{x}_1, \mathbf{x}_2, \mathbf{t}, \mathbf{v}\big) \in \Upsilon_+^b \times \Upsilon_-^b \times \Upsilon_+^x \times \Upsilon_+^x \times \Upsilon_S \times \mathbb{I}^{W-1}$.
- *Type III vicinity :* $\big(\mathbf{b}_1, \mathbf{b}_2, \mathbf{x}_1, \mathbf{x}_2, \mathbf{t}, \mathbf{v}\big) \in \Upsilon_+^b \times \Upsilon_-^b \times \Upsilon_-^x \times \Upsilon_-^x \times \Upsilon_S \times \mathbb{I}^{W-1}$.

In the following discussion, the parameter ε_0 in Θ is allowed to be different from line to line. However, given ε_1 in (1.7), we shall always choose ε_2 in (1.15) and ε_0 in (5.33) according to the rule

$$C\varepsilon_2 \le \varepsilon_0 \le \frac{\varepsilon_1}{C} \tag{5.37}$$

for some sufficiently large $C > 0$. Consequently, by Assumption 1.13 we have

$$N(\log N)^{-10} \ge M = M^{1-4\varepsilon_0}M^{4\varepsilon_0} \ge W^{(4+2\gamma+\varepsilon_1)(1-4\varepsilon_0)}M^{4\varepsilon_0} \gg W^{(4+2\gamma+4\varepsilon_0)}M^{4\varepsilon_0} = W^{2\gamma}\Theta^4. \tag{5.38}$$

To prove Theorem 1.14, we split the task into three steps. The first step is to exclude the integral outside the vicinities. Specifically, we will show the following lemma.

Lemma 5.6. *Under Assumptions 1.1 and 1.13, we have,*

$$\begin{aligned}\mathcal{I}\big(\Gamma^W, \bar\Gamma^W, \Sigma^W, \Sigma^W, \mathbb{R}_+^{W-1}, \mathbb{I}^{W-1}\big) = 2^W\mathcal{I}\big(\Upsilon_+^b, \Upsilon_-^b, \Upsilon_+^x, \Upsilon_-^x, \Upsilon_S, \Upsilon_S\big)\\ + \mathcal{I}\big(\Upsilon_+^b, \Upsilon_-^b, \Upsilon_+^x, \Upsilon_+^x, \Upsilon_S, \mathbb{I}^{W-1}\big)\\ + \mathcal{I}\big(\Upsilon_+^b, \Upsilon_-^b, \Upsilon_-^x, \Upsilon_-^x, \Upsilon_S, \mathbb{I}^{W-1}\big) + O(e^{-\Theta}).\end{aligned} \tag{5.39}$$

Remark 5.7. *The first three terms on the r.h.s. of (5.39) correspond to the integrals over vicinities of the Type I, II, and III saddle points, respectively. Note that for the first term, we have used the argument in Section 4.1, namely, the total contribution of the integral over the Type I vicinity is 2^W times that over the Type I' vicinity.*

The second step, is to estimate the integral over the Type I vicinity. We have the following lemma.

Lemma 5.8. *Under Assumptions 1.1 and 1.13, there exists some positive constant C_0 uniform in n and some positive number $N_0 = N_0(n)$ such that for all $N \ge N_0$,*

$$2^W\big|\mathcal{I}(\Upsilon_+^b, \Upsilon_-^b, \Upsilon_+^x, \Upsilon_-^x, \Upsilon_S, \Upsilon_S)\big| \le \frac{N^{C_0}}{(N\eta)^n}. \tag{5.40}$$

The last step is to show that the integral over the Type II and III vicinities are also negligible.

Lemma 5.9. *Under Assumptions 1.1 and 1.13, there exists some positive constant c such that,*

$$\mathcal{I}\big(\Upsilon_+^b, \Upsilon_-^b, \Upsilon_+^x, \Upsilon_+^x, \Upsilon_S, \mathbb{I}^{W-1}\big) = O(e^{-cW}), \qquad \mathcal{I}\big(\Upsilon_+^b, \Upsilon_-^b, \Upsilon_-^x, \Upsilon_-^x, \Upsilon_S, \mathbb{I}^{W-1}\big) = O(e^{-cW}).$$

Therefore, the remaining task is to prove Lemmas 5.1, 5.6, 5.8 and 5.9. For the convenience of the reader, we outline the organization of the subsequent part as follows.

At first, the proofs of Lemmas 5.1 and 5.6 require a discussion on the bound of the integrand, especially on the term $\mathsf{A}(\cdot)$, which contains the integral over all the Grassmann variables. To this end, we perform a crude analysis for the function $\mathsf{A}(\cdot)$ in Section 6 in advance, with which we are able to prove Lemmas 5.1 and 5.6 in Section 7.

Then, we can restrict ourselves to the integral over the vicinities, i.e., prove Lemmas 5.8 and 5.9. It will be shown that in the vicinity $(\mathbf{b}_1, \mathbf{b}_2, \mathbf{t}) \in \Upsilon_+^b \times \Upsilon_-^b \times \Upsilon_S$, the factor $\exp\{-M\mathring{L}(\hat B, T)\}$ is approximately the product of a complex Gaussian measure of the $\hat B$-variables and a real Gaussian measure of the $\mathbf{t}$-variables. Here, by "complex Gaussian measure" we mean a function of the form $\exp\{-\mathbf{u}'\mathbb{A}\mathbf{u}\}$, where $\mathbf{u}$ is a real vector, while $\mathbb{A}$ is a complex matrix with positive-definite Hermitian part. In order to estimate the integral against this Gaussian measure (in an approximate sense), we shall get rid of the $o(1)$ term in the integral of the form

$$\int \mathrm{d}\mathbf{u}\ \exp\big\{-\mathbf{u}'\mathbb{A}\mathbf{u} + o(1)\big\}\,\mathfrak{f}(\mathbf{u}), \tag{5.41}$$

for some function $\mathfrak{f}$, which however cannot be done directly, owing to the fact that $\mathbb{A}$ is complex. In our case, this problem can be solved by further deforming the contours of the $\hat B$-variables, following the

steepest descent paths exactly in the vicinity. By doing this, we can get a real Gaussian measure, thus the remainder terms can be easily controlled when integrate against this measure. The situation for $\exp\{-M\mathring{K}(\hat{X},V)\}$ is a little bit more complicated due to different types of the saddle points. However, in the Type I vicinity, we can do the same thing. Hence, in Section 8, we will analyze the Gaussian measure (in an approximate sense) $\exp\{-M(\mathring{K}(\hat{X},V)+\mathring{L}(\hat{B},T))\}$, especially, we will further deform the contours of $\hat{X}$ and $\hat{B}$-variables in the vicinities, whereby we can prove Lemmas 5.8 in Section 9. In the Type II and III vicinities, we will bound $\exp\{-M\mathring{K}(\hat{X},V)\}$ by its absolute value directly. It turns out to be enough for our proof of Lemma 5.9, which is given in Section 10.

6. Crude bound on $\mathsf{A}(\hat{X},\hat{B},V,T)$

In this section, we provide a bound on the function $\mathsf{A}(\cdot)$ in terms of the $\hat{B},T$-variables, which holds on all the domains under discussion in the sequel. Here, by *crude bound* we mean a bound of order $\exp\{O(WN^{\varepsilon_2})\}$), which will be specified in Lemma 6.1 below. By the definition in (3.31), we see that $\mathsf{A}(\cdot)$ is an integral of the product of $\mathcal{Q}(\cdot)$, $\mathcal{P}(\cdot)$ and $\mathcal{F}(\cdot)$. We will mainly treat $\mathcal{Q}(\cdot)$ as a function of $\boldsymbol{\omega}^{[1]},\boldsymbol{\xi}^{[1]}$-variables, treat $\mathcal{P}(\cdot)$ as a function of Ω,Ξ-variables, and treat $\mathcal{F}(\cdot)$ as a function of $X^{[1]},\mathbf{y}^{[1]},\mathbf{w}^{[1]},P_1,Q_1$-variables. However, in the function $\mathcal{Q}(\cdot)$, we actually have every argument mentioned above. Hence, we perform the integral over $\boldsymbol{\omega}^{[1]}$-variables and $\boldsymbol{\xi}^{[1]}$-variables for $\mathcal{Q}(\cdot)$ at first. The resulting function $\mathsf{Q}(\cdot)$ turns out to be a polynomial of the remaining arguments. As mentioned in Section 4.2, a typical procedure we will adopt is to ignore $\mathsf{Q}(\cdot)$ at first, then estimate the integrals of $\mathcal{P}(\cdot)$ and $\mathcal{F}(\cdot)$, which are denoted by $\mathsf{P}(\cdot)$ and $\mathsf{F}(\cdot)$, respectively (see (4.4) and (4.5)), finally, we make necessary comment on how to modify the bounding scheme to take $\mathsf{Q}(\cdot)$ into account, whereby we can get the desired bound for $\mathsf{A}(\cdot)$.

For the sake of simplicity, from now on, we will use the notation

$$\begin{aligned}&\omega_{j,1}:=\omega_{j,11},\quad \omega_{j,2}:=\omega_{j,12},\quad \omega_{j,3}:=\omega_{j,21},\quad \omega_{j,4}:=\omega_{j,22},\\&\xi_{j,1}:=\xi_{j,11},\quad \xi_{j,2}:=\xi_{j,21},\quad \xi_{j,3}:=\xi_{j,12},\quad \xi_{j,4}:=\xi_{j,22}.\end{aligned}\tag{6.1}$$

Moreover, we introduce the domains

$$\widehat{\Sigma}:=\Big\{re^{\mathbf{i}\vartheta}:|r-1|\le\frac{1}{10},\vartheta\in\mathbb{L}\Big\},$$

$$\mathbb{K}\equiv\mathbb{K}(E):=\begin{cases}\Big\{\omega\in\mathbb{C}:0\le\arg\omega\le\frac{\arg a_+}{2}+\frac{\pi}{8}\Big\}, & \text{if}\quad E\ge 0,\\ \Big\{\omega\in\mathbb{C}:\frac{\arg a_+}{2}-\frac{\pi}{8}\le\arg\omega\le 0\Big\}, & \text{if}\quad E<0.\end{cases}\tag{6.2}$$

By the assumption that $|E|\le\sqrt{2}-\kappa$ in (1.15), it is easy to see that $|\arg\omega|\le\pi/4-c$ for all $\omega\in\mathbb{K}\cup\bar{\mathbb{K}}$, where c is some positive constant depending on κ. Our aim is to show the following lemma.

Lemma 6.1. *Suppose that $\mathbf{b}_1,\mathbf{b}_2,\mathbf{x}_1,\mathbf{x}_2\in\mathbb{K}^W\times\bar{\mathbb{K}}^W\times\widehat{\Sigma}^W\times\widehat{\Sigma}^W$. Under the assumption of Theorem 1.14, we have*

$$|\mathsf{A}(\hat{X},\hat{B},V,T)|\le e^{O(WN^{\varepsilon_2})}\prod_{j=1}^{W}\Big(r_{j,1}^{-1}+r_{j,2}^{-1}+t_j+1\Big)^C:=e^{O(WN^{\varepsilon_2})}\mathfrak{p}(\mathbf{r}^{-1},\mathbf{t}).$$

Remark 6.2. *Obviously, using the terminology introduced at the end of Section 3.6, we have*

$$\mathfrak{p}(\mathbf{r}^{-1},\mathbf{t})\in\mathfrak{Q}\Big(\{r_{j,1}^{-1},r_{j,2}^{-1},t_j\}_{j=1}^{W};\kappa_1,\kappa_2,\kappa_3\Big),\quad \kappa_1=e^{O(W)},\quad \kappa_2,\kappa_3=O(1).\tag{6.3}$$

6.1. **Integral of $\mathcal{Q}$.** In this section, we investigate the function

$$\mathtt{Q}(\Omega,\Xi,P_1,Q_1,X^{[1]},\mathbf{y}^{[1]},\mathbf{w}^{[1]}):=\int \mathrm{d}\boldsymbol{\omega}^{[1]}\mathrm{d}\boldsymbol{\xi}^{[1]}\ \mathcal{Q}(\Omega,\Xi,\boldsymbol{\omega}^{[1]},\boldsymbol{\xi}^{[1]},P_1,Q_1,X^{[1]},\mathbf{y}^{[1]},\mathbf{w}^{[1]}). \tag{6.4}$$

Recall $\mathfrak{Q}_{\mathrm{deg}}(\mathbf{a};\kappa_1,\kappa_2,\kappa_3)$ defined at the end of Section 3.6, the parameterization in (3.15) and (3.24) and the notation introduced in (6.1). We shall show the following lemma.

Lemma 6.3. *If we regard σ, $v_p^{[1]}$, $v_q^{[1]}$, P_1 and $(X^{[1]})^{-1}$-entries as fixed parameters, we have*

$$\mathtt{Q}(\cdot)\in\mathfrak{Q}_{deg}(\mathfrak{S};\kappa_1,\kappa_2,\kappa_3),\qquad \kappa_1=W^{O(1)},\quad \kappa_2,\kappa_3=O(1), \tag{6.5}$$

where $\mathfrak{S}$ is the set of variables defined by

$$\mathfrak{S}:=\Big\{t,s,(y_k^{[1]})^{-1},e^{\mathbf{i}\sigma_k^{[1]}},e^{-\mathbf{i}\sigma_k^{[1]}},\frac{\omega_{i,a}\xi_{j,b}}{M}\Big\}_{\substack{i,j=1,\dots,W;\\ k=p,q;a,b=1,\dots,4}}.$$

Proof. Note that $\mathcal{Q}(\cdot)$ can be regarded as a function of the Grassmann variables in $\boldsymbol{\omega}^{[1]}$ and $\boldsymbol{\xi}^{[1]}$. Hence, by the definition in (3.29), it is a polynomial of these variables with bounded degree. To have a closer look on this polynomial, we start with

$$\exp\Big\{-\frac{1}{\sqrt{M}}\sum_j\tilde{\mathfrak{s}}_{jk}TrP_1^*\Omega_jQ_1\mathbf{w}_k^{[1]}\boldsymbol{\xi}_k^{[1]}J\Big\},\qquad k=p,q. \tag{6.6}$$

Observe that, in the exponent of (6.6), one $\boldsymbol{\xi}_k^{[1]}$-variable must be accompanied by one Ω-variable. In addition, we combine the factor $1/\sqrt{M}$ with Ω_j's. Then, by Taylor expansion with respect to $\boldsymbol{\xi}_k^{[1]}$-variables, it is easy to find

$$\exp\Big\{-\frac{1}{\sqrt{M}}\sum_j\tilde{\mathfrak{s}}_{jk}TrP_1^*\Omega_jQ_1\mathbf{w}_k^{[1]}\boldsymbol{\xi}_k^{[1]}J\Big\}\in\mathfrak{Q}_{\mathrm{deg}}(\mathfrak{S}_{1,k};\kappa_1,\kappa_2,\kappa_3),\quad \kappa_1=W^{O(1)},\quad \kappa_2,\kappa_3=O(1), \tag{6.7}$$

where

$$\mathfrak{S}_{1,k}:=\Big\{t,s,e^{\mathbf{i}\sigma_k^{[1]}},\frac{\omega_{j,a}\xi_{k,b}^{[1]}}{\sqrt{M}}\Big\}_{\substack{j=1,\dots,W;\\ a=1,\dots,4;b=1,2}},\qquad k=p,q.$$

Analogously, we can show that for $k=p,q$

$$\exp\Big\{-\frac{1}{\sqrt{M}}\sum_j\tilde{\mathfrak{s}}_{kj}Tr\boldsymbol{\omega}_k^{[1]}(\mathbf{w}_k^{[1]})^*JQ_1^{-1}\Xi_jP_1\Big\}\in\mathfrak{Q}_{\mathrm{deg}}(\mathfrak{S}_{2,k};\kappa_1,\kappa_2,\kappa_3),\qquad \kappa_1=W^{O(1)},\quad \kappa_2,\kappa_3=O(1), \tag{6.8}$$

where

$$\mathfrak{S}_{2,k}:=\Big\{t,s,e^{-\mathbf{i}\sigma_k^{[1]}},\frac{\xi_{j,a}\omega_{k,b}^{[1]}}{\sqrt{M}}\Big\}_{\substack{j=1,\dots,W;\\ a=1,\dots,4;b=1,2}},\quad k=p,q.$$

Here we have used the fact that Q_1^{-1}-entries are the same as Q_1-entries, up to a sign. In a similar manner, one can show that for $k,\ell=p,q$,

$$\exp\Big\{-\frac{1}{M}\tilde{\mathfrak{s}}_{k\ell}Tr\boldsymbol{\omega}_k^{[1]}(\mathbf{w}_k^{[1]})^*J(\mathbf{w}_\ell^{[1]})\boldsymbol{\xi}_\ell^{[1]}J\Big\}\in\mathfrak{Q}_{\mathrm{deg}}(\mathfrak{S}_{3,k,\ell};\kappa_1,\kappa_2,\kappa_3),\quad \kappa_1,\kappa_2,\kappa_3=O(1), \tag{6.9}$$

and

$$\Big(1-(y_k^{[1]})^{-1}\boldsymbol{\xi}_k^{[1]}(X_k^{[1]})^{-1}\boldsymbol{\omega}_k^{[1]}\Big)^2\in\mathfrak{Q}_{\mathrm{deg}}(\mathfrak{S}_{4,k};\kappa_1,\kappa_2,\kappa_3),\quad \kappa_1,\kappa_2,\kappa_3=O(1), \tag{6.10}$$

where

$$\mathfrak{S}_{3,k,\ell}:=\Big\{e^{\mathbf{i}\sigma_k^{[1]}},e^{-\mathbf{i}\sigma_\ell^{[1]}},\omega_{k,a}^{[1]}\xi_{\ell,b}^{[1]}\Big\}_{a,b=1,2},\quad \mathfrak{S}_{4,k}:=\Big\{(y_k^{[1]})^{-1},\omega_{k,a}^{[1]}\xi_{\ell,b}^{[1]}\Big\}_{a,b=1,2}.$$

Hence, by (6.7)-(6.10) and (3.29), we see that

$$\mathcal{Q}(\cdot) \in \mathfrak{Q}_{\deg}(\mathfrak{S}_5; \kappa_1, \kappa_2, \kappa_3), \quad \kappa_1 = W^{O(1)}, \quad \kappa_2, \kappa_3 = O(1), \tag{6.11}$$

where

$$\mathfrak{S}_5 := \Big\{ t, s, (y_k^{[1]})^{-1}, e^{\mathbf{i}\sigma_k^{[1]}}, e^{-\mathbf{i}\sigma_k^{[1]}}, \frac{\omega_{j,r}\xi_{k,b}^{[1]}}{\sqrt{M}}, \frac{\xi_{j,r}\omega_{k,b}^{[1]}}{\sqrt{M}}, \omega_{k,a}^{[1]}\xi_{\ell,b}^{[1]} \Big\}_{\substack{j=1,\dots,W; k=p,q; \\ r=1,\dots,4; a,b=1,2}}.$$

By the definition in (6.4), $\mathsf{Q}(\cdot)$ is the integral of $\mathcal{Q}(\cdot)$ over the $\boldsymbol{\omega}^{[1]}$ and $\boldsymbol{\xi}^{[1]}$-variables. Now, we regard all the other variables in $\mathfrak{S}_5$, except $\boldsymbol{\omega}^{[1]}$ and $\boldsymbol{\xi}^{[1]}$-variables, as parameters. By the definition of Grassmann integral, we know that only the coefficient of the highest order term $\prod_{k=p,q}\prod_{a=1,2}\omega_{k,a}^{[1]}\xi_{k,a}^{[1]}$ in $\mathcal{Q}(\cdot)$ survives after integrating $\boldsymbol{\omega}^{[1]}$ and $\boldsymbol{\xi}^{[1]}$-variables out. Then, it is easy to see (6.5) from (6.11), completing the proof. □

6.2. **Integral of $\mathcal{P}$.** In this subsection, we temporarily ignore the Ω and Ξ-variables from $\mathsf{Q}(\cdot)$, and estimate $\mathsf{P}(\hat{X}, \hat{B}, V, T)$ defined in (4.4). Recalling $r_{j,1}$ and $r_{j,2}$ defined in (5.12), we can formulate our estimate as follows.

Lemma 6.4. *Suppose that the assumptions in Lemma 6.1 hold. We have*

$$|\mathsf{P}(\hat{X}, \hat{B}, V, T)| \le e^{O(W)} \prod_{j=1}^{W} \big(r_{j,1}^{-1} + r_{j,2}^{-1} + t_j + 1\big)^{O(1)}. \tag{6.12}$$

Proof. We start with one factor from $\mathcal{P}(\cdot)$ (see (3.28)), namely

$$\begin{aligned} \boldsymbol{\varpi}_j &:= \frac{1}{\det{}^M\big(1 + M^{-1}V_j^*\hat{X}_j^{-1}V_j\Omega_j T_j^{-1}\hat{B}_j^{-1}T_j\Xi_j\big)} \\ &= \exp\Big\{ -M\log\det\big(1 + M^{-1}V_j^*\hat{X}_j^{-1}V_j\Omega_j T_j^{-1}\hat{B}_j^{-1}T_j\Xi_j\big)\Big\} \\ &= 1 + \sum_{\ell=1}^{4} \frac{1}{M^{\ell-1}}\mathfrak{p}_\ell(\hat{X}_j, \hat{B}_j, V_j, T_j, \Omega_j, \Xi_j). \end{aligned} \tag{6.13}$$

Here $\mathfrak{p}_\ell(\cdot)$ is a polynomial in $\hat{X}_j^{-1}$, $\hat{B}_j^{-1}$, V_j, T_j, Ω_j and Ξ_j-entries with bounded degree and bounded coefficients. Here we used the fact that V_j^* and T_j^{-1}-entries are the same as V_j and T_j-entries, respectively, up to a sign. Moreover, if we regard $\mathfrak{p}_\ell(\cdot)$ as a polynomial of Ω_j and Ξ_j-entries, it is homogeneous, with degree 2ℓ, and the total degree for Ω_j-variables is ℓ, thus that for Ξ_j-entries is also ℓ. More specifically, we can write

$$\mathfrak{p}_\ell(\hat{X}_j, \hat{B}_j, V_j, T_j, \Omega_j, \Xi_j) = \sum_{\substack{\alpha_1,\dots,\alpha_\ell, \\ \beta_1,\dots,\beta_\ell=1}}^{4} \mathfrak{p}_{\ell,\boldsymbol{\alpha},\boldsymbol{\beta}}(\hat{X}_j, \hat{B}_j, V_j, T_j) \prod_{i=1}^{\ell} \omega_{j,\alpha_i}\xi_{j,\beta_i},$$

where we used the notation in (6.1) and denoted $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_\ell)$ and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_\ell)$. It is easy to verify that $\boldsymbol{\varpi}_j$ is of the form (6.13) by taking Taylor expansion with respect to the Grassmann variables. The expansion in (6.13) terminates at $\ell = 4$, owing to the fact that there are totally 8 Grassmann variables from Ω_j and Ξ_j. In addition, it is also easy to check that $\mathfrak{p}_{\ell,\boldsymbol{\alpha},\boldsymbol{\beta}}(\cdot)$ is a polynomial of $\hat{X}_j^{-1}$, $\hat{B}_j^{-1}$, V_j, T_j-entries with bounded degree and bounded coefficients, which implies that there exist two positive constants C_1 and C_2, such that

$$|\mathfrak{p}_{\ell,\boldsymbol{\alpha},\boldsymbol{\beta}}(\cdot)| \le C_1\big(r_{j,1}^{-1} + r_{j,2}^{-1} + t_j + 1\big)^{C_2} \tag{6.14}$$

uniformly in ℓ, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Here we used the fact that $\hat{X}_j^{-1}$ and V_j-entries are all bounded and T_j-entries are bounded by $1 + t_j$.

Now, we go back to the definition of $\mathcal{P}(\cdot)$ in (3.28) and study the last factor. Similarly to the discussion above, it is easy to see that for $k = p$ or q,

$$\hat{\boldsymbol{\varpi}}_k := \frac{\det\left(V_k^* \hat{X}_k V_k + M^{-1}\Omega_k T_k^{-1} \hat{B}_k^{-1} T_k \Xi_k\right)}{\det \hat{B}_k}$$
$$= \hat{\mathfrak{p}}_0(\hat{X}_k, \hat{B}_k) + \sum_{\ell=1}^{4} \sum_{\boldsymbol{\alpha},\boldsymbol{\beta}} \hat{\mathfrak{p}}_{\ell,\boldsymbol{\alpha},\boldsymbol{\beta}}(\hat{X}_k, \hat{B}_k, V_k, T_k) \prod_{i=1}^{\ell} \omega_{k,\alpha_{k,i}} \xi_{k,\beta_{k,i}}, \tag{6.15}$$

where $\hat{\mathfrak{p}}_0(\cdot) = \det \hat{X}_k / \det \hat{B}_k$ and $\hat{\mathfrak{p}}_{\ell,\boldsymbol{\alpha},\boldsymbol{\beta}}(\cdot)$'s are some polynomials of $\hat{X}_k$, $\hat{B}_k^{-1}$, V_k, T_k-entries with bounded degree and bounded coefficients. Similarly, we have

$$|\hat{\mathfrak{p}}_0(\cdot)|, |\hat{\mathfrak{p}}_{\ell,\boldsymbol{\alpha},\boldsymbol{\beta}}(\cdot)| \le C_1 (r_{k,1}^{-1} + r_{k,2}^{-1} + t_k + 1)^{C_2} \tag{6.16}$$

for some positive constants C_1 and C_2.

According to the definitions in (6.13) and (6.15), we can rewrite (3.28) as

$$\mathcal{P}(\Omega, \Xi, \hat{X}, \hat{B}, V, T) = \exp\Big\{ -\sum_{j,k} \tilde{\mathfrak{s}}_{jk} Tr \Omega_j \Xi_k \Big\} \prod_{j=1}^{W} \boldsymbol{\varpi}_j \prod_{k=p,q} \hat{\boldsymbol{\varpi}}_k. \tag{6.17}$$

In light of the discussion above, $\prod_{j=1}^{W} \boldsymbol{\varpi}_j \prod_{k=p,q} \hat{\boldsymbol{\varpi}}_k$ is a polynomial of $\hat{X}^{-1}$, $\hat{B}^{-1}$, V, T, Ω and Ξ-entries, in which each monomial is of the form

$$\mathfrak{q}_{\vec{\ell},\vec{\boldsymbol{\alpha}},\vec{\boldsymbol{\beta}}}(\hat{X}^{-1}, \hat{B}^{-1}, V, T) \prod_{j=1}^{W} \prod_{i=1}^{\ell_j} \omega_{j,\alpha_{j,i}} \xi_{j,\beta_{j,i}}, \tag{6.18}$$

where we used the notation

$$\vec{\ell} = (\ell_1, \dots, \ell_W), \quad \vec{\boldsymbol{\alpha}} = (\boldsymbol{\alpha}_1, \dots, \boldsymbol{\alpha}_W), \quad \vec{\boldsymbol{\beta}} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_W),$$
$$\boldsymbol{\alpha}_j = (\alpha_{j,1}, \dots, \alpha_{j,\ell_j}), \quad \boldsymbol{\beta}_j = (\beta_{j,1}, \dots, \beta_{j,\ell_j}),$$

and $\mathfrak{q}_{\vec{\ell},\vec{\boldsymbol{\alpha}},\vec{\boldsymbol{\beta}}}(\cdot)$ is a polynomial of $\hat{X}$, $\hat{X}^{-1}$, $\hat{B}^{-1}$, V and T-entries. Moreover, all the entries of $\vec{\ell}$, $\vec{\boldsymbol{\alpha}}$ and $\vec{\boldsymbol{\beta}}$ are bounded by 4. By (6.14) and (6.16), we have

$$|\mathfrak{q}_{\vec{\ell},\vec{\boldsymbol{\alpha}},\vec{\boldsymbol{\beta}}}(\hat{X}^{-1}, \hat{B}^{-1}, V, T)| \le e^{O(W)} \prod_{j=1}^{W} \left(r_{j,1}^{-1} + r_{j,2}^{-1} + t_j + 1\right)^{C}. \tag{6.19}$$

In addition, it is easy to see that the number of the summands of the form (6.18) in $\prod_{j=1}^{W} \boldsymbol{\varpi}_j \prod_{k=p,q} \hat{\boldsymbol{\varpi}}_k$ is bounded by $e^{O(W)}$.

Define the vectors

$$\vec{\Omega} := (\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \boldsymbol{\omega}_3, \boldsymbol{\omega}_4), \qquad \vec{\Xi} := (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \boldsymbol{\xi}_3, \boldsymbol{\xi}_4), \tag{6.20}$$

where

$$\boldsymbol{\omega}_\alpha = (\omega_{1,\alpha}, \dots, \omega_{W,\alpha}), \quad \boldsymbol{\xi}_\alpha = (\xi_{1,\alpha}, \dots, \xi_{W,\alpha}), \quad \alpha = 1, 2, 3, 4.$$

Here we used the notation (6.1). In addition, we introduce the matrix

$$\tilde{\mathbb{H}} = \tilde{S} \oplus \tilde{S} \oplus \tilde{S} \oplus \tilde{S}.$$

It is easy to check

$$\sum_{j,k} \tilde{\mathfrak{s}}_{jk} Tr \Omega_j \Xi_k = \vec{\Omega} \tilde{\mathbb{H}} \vec{\Xi}'.$$

By using the Gaussian integral formula for the Grassmann variables (3.2), we see that for each $\vec{\ell}$, $\vec{\boldsymbol{\alpha}}$ and $\vec{\boldsymbol{\beta}}$, we have

$$\Big|\int \mathrm{d}\Omega\mathrm{d}\Xi\cdot\exp\Big\{-\sum_{j,k}\tilde{\mathfrak{s}}_{jk}Tr\Omega_j\Xi_k\Big\}\cdot\prod_{j=1}^{W}\prod_{i=1}^{\ell_j}\omega_{j,\alpha_{j,i}}\xi_{j,\beta_{j,i}}\Big|\leq|\det\tilde{\mathbb{H}}^{(\mathsf{I}|\mathsf{J})}|, \tag{6.21}$$

for some index sets I and J with $|\mathsf{I}|=|\mathsf{J}|$. By Assumption 1.1 (i) and (ii), we see that the 2-norm of each row of $\tilde{S}$ is $O(1)$. Consequently, by using Hadamard's inequality, we have

$$|\det\tilde{\mathbb{H}}^{(\mathsf{I}|\mathsf{J})}|=e^{O(W)}. \tag{6.22}$$

Therefore, (6.17)-(6.22) and the bound $e^{O(W)}$ for the total number of summands of the form (6.18) in $\prod_{j=1}^{W}\boldsymbol{\varpi}_j\prod_{k=p,q}\hat{\boldsymbol{\varpi}}_k$ imply that

$$|\mathsf{P}(\hat{X},\hat{B},V,T)|\leq e^{O(W)}\prod_{j=1}^{W}\big(r_{j,1}^{-1}+r_{j,2}^{-1}+t_j+1\big)^{O(1)}.$$

Thus we completed the proof. □

6.3. **Integral of $\mathcal{F}$.** In this subsection, we also temporarily ignore the $X^{[1]},\mathbf{y}^{[1]},\mathbf{w}^{[1]},P_1,Q_1$-variables from $\mathsf{Q}(\cdot)$, and estimate $\mathsf{F}(\hat{X},\hat{B},V,T)$ defined in (4.5). We have the following lemma.

Lemma 6.5. *Suppose that the assumptions in Lemma 6.1 hold. We have*

$$|\mathsf{F}(\hat{X},\hat{B},V,T)|\leq e^{O(WN^{\varepsilon_2})}\prod_{k=p,q}\big(r_{k,1}^{-1}+r_{k,2}^{-1}+t_k+1\big)^{O(1)}.$$

Proof. Recalling the decomposition of $\mathcal{F}(\cdot)$ in (3.18) together with the parameterization in (3.27), we will study the integrals

$$\mathbb{G}(\hat{B},T):=\int\mathrm{d}\nu(Q_1)\mathrm{d}\mathbf{y}^{[1]}\mathrm{d}\mathbf{w}^{[1]}\,g(Q_1,T,\hat{B},\mathbf{y}^{[1]},\mathbf{w}^{[1]}), \tag{6.23}$$

$$\mathbb{F}(\hat{X},V):=\int\mathrm{d}\mu(P_1)\mathrm{d}X^{[1]}\,f(P_1,V,\hat{X},X^{[1]}) \tag{6.24}$$

separately. Recalling the convention at the end of Section 3, we use $f(\cdot)$ and $g(\cdot)$ to represent the integrands above. One can refer to (3.19) and (3.20) for the definition.

From the assumption $\eta\leq M^{-1}N^{\varepsilon_2}$, it is easy to see

$$|\mathbb{F}(\hat{X},V)|\leq e^{O(WN^{\varepsilon_2})}, \tag{6.25}$$

since $P_1,V,\hat{X},X^{[1]}$-variables are all bounded and $|\det X_p^{[1]}|,|\det X_q^{[1]}|\sim 1$ when $\mathbf{x}_1,\mathbf{x}_2\in\widehat{\Sigma}$ defined in (6.2).

For $\mathbb{G}(\hat{B},T)$, we use the facts

$$\begin{gathered}\mathsf{Re}(TrB_jY_k^{[1]}J)\geq 0,\quad \mathsf{Re}(\mathbf{i}TrY_k^{[1]}JZ)=-\eta TrY_k^{[1]}\leq 0,\quad TrY_k^{[1]}JY_\ell^{[1]}J\geq 0,\quad k,\ell=p,q,\\ \big|(\mathbf{w}_q^{[1]}(\mathbf{w}_q^{[1]})^*)_{12}\big|\leq 1,\quad \big|(\mathbf{w}_p^{[1]}(\mathbf{w}_p^{[1]})^*)_{21}\big|\leq 1,\end{gathered} \tag{6.26}$$

to estimate trivially several terms, whereby we can get the bound

$$|g(\cdot)|\leq\exp\Big\{-M\eta\sum_{j=1}^{W}Tr\mathsf{Re}(B_j)J\Big\}\prod_{k=p,q}(y_k^{[1]})^{n+3}\exp\Big\{-\tilde{\mathfrak{s}}_{kk}Tr\mathsf{Re}(B_k)Y_k^{[1]}J\Big\}. \tag{6.27}$$

Here $\mathsf{Re}(B_j)=Q_1^{-1}T_j^{-1}\mathsf{Re}(\hat{B}_j)T_jQ_1$. Hence, integrating $y_p^{[1]}$ and $y_q^{[1]}$ out yields

$$\int_{\mathbb{R}_+^2}\mathrm{d}y_p^{[1]}\mathrm{d}y_q^{[1]}|g(Q_1,T,\hat{B},\mathbf{y}^{[1]},\mathbf{w}^{[1]})|\leq C\frac{\exp\Big\{-M\eta\sum_{j=1}^{W}Tr\mathsf{Re}(B_j)J\Big\}}{\prod_{k=p,q}\Big((\mathbf{w}_k^{[1]})^*J\mathsf{Re}(B_k)\mathbf{w}_k^{[1]}\Big)^{C_1}}, \tag{6.28}$$

for some positive constants C and C_1 depending on n, where we used the elementary facts that $\tilde{\mathfrak{s}}_{kk} \geq c$ for some positive constant c and

$$Tr\mathsf{Re}(B_j)Y_k^{[1]}J = y_k^{[1]}(\mathbf{w}_k^{[1]})^* J\mathsf{Re}(B_j)\mathbf{w}_k^{[1]}, \quad k = p, q, \quad j = 1, \ldots, W. \tag{6.29}$$

Now, note that

$$(\mathbf{w}_k^{[1]})^* J\mathsf{Re}B_j\mathbf{w}_k^{[1]} \geq \lambda_1(J\mathsf{Re}B_j), \quad k = p, q, \quad j = 1, \ldots, W. \tag{6.30}$$

In addition, it is also easy to see $\lambda_1(T_j) = s_j - t_j$ and $\lambda_1(Q_1) = s - t$, according to the definitions in (3.24). Now, by the fact $JA^{-1} = AJ$ for any $A \in \mathring{U}(1,1)$, we have

$$J\mathsf{Re}B_j = Q_1T_j\mathrm{diag}\big(\mathsf{Re}b_{j,1}, \mathsf{Re}b_{j,2}\big)T_jQ_1. \tag{6.31}$$

Consequently, we can get

$$\lambda_1(J\mathsf{Re}B_j) \geq (s_j - t_j)^2(s-t)^2\min\{\mathsf{Re}b_{j,1}, \mathsf{Re}b_{j,2}\} = \frac{\min\{\mathsf{Re}b_{j,1}, \mathsf{Re}b_{j,2}\}}{(s_j + t_j)^2(s+t)^2}, \tag{6.32}$$

by recalling the facts $s^2 - t^2 = 1$ and $s_j^2 - t_j^2 = 1$. Therefore, combining (6.28), (6.30) and (6.32), we have

$$\int_{\mathbb{R}_+^2} \mathrm{d}y_p^{[1]}\mathrm{d}y_q^{[1]}|g(\cdot)| \leq C(s+t)^{4C_1}\exp\Big\{-M\eta\sum_{j=1}^W Tr(\mathsf{Re}B_j)J\Big\}\prod_{k=p,q}\frac{(s_k+t_k)^{2C_1}}{\big(\min\{\mathsf{Re}b_{k,1}, \mathsf{Re}b_{k,2}\}\big)^{C_1}}. \tag{6.33}$$

Now, what remains is to estimate the exponential function in (6.33). By elementary calculation from (6.31) we obtain

$$Tr(\mathsf{Re}B_j)J \geq \big(\mathsf{Re}b_{j,1} + \mathsf{Re}b_{j,2}\big)\big((s_j^2 + t_j^2)(s^2 + t^2) - 4sts_jt_j\big).$$

Observe that

$$(s_j^2 + t_j^2)(s^2 + t^2) - 4sts_jt_j = \frac{(s_j^2 + t_j^2)^2(s^2 + t^2)^2 - 16(sts_jt_j)^2}{(s_j^2 + t_j^2)(s^2 + t^2) + 4sts_jt_j} \geq \frac{4t^4 + 4t^2 + 4t_j^4 + 4t_j^2 + 1}{2(1 + 2t_j^2)(1 + 2t^2)} \geq \frac{1 + 2t^2}{2(1 + 2t_j^2)}.$$

It implies that

$$\begin{aligned}\exp\Big\{-M\eta\sum_{j=1}^W Tr(\mathsf{Re}B_j)J\Big\} &\leq \exp\Big\{-2M\eta\sum_{j=1}^W\frac{\mathsf{Re}b_{j,1} + \mathsf{Re}b_{j,2}}{1 + 2t_j^2}(1 + 2t^2)\Big\} \\ &\leq \exp\Big\{-cM\eta\sum_{j=1}^W\frac{r_{j,1} + r_{j,2}}{1 + 2t_j^2}(1 + 2t^2)\Big\},\end{aligned} \tag{6.34}$$

for some positive constant c, where in the last step we used the fact

$$\mathsf{Re}b_{j,\alpha} \geq cr_{j,\alpha}, \quad \forall\, j = 1, \ldots, W, \alpha = 1, 2, \tag{6.35}$$

for some positive constant c, in light of the assumption $|E| \leq \sqrt{2} - \kappa$ and the definition of $\mathbb{K}$ in (6.2). Plugging (6.34) into (6.33), estimating $(s+t)^2 \leq 2(1+2t^2)$, and integrating t out, we can crudely bound

$$\int_{\mathbb{R}_+^2}\mathrm{d}y_p^{[1]}\mathrm{d}y_q^{[1]}\int_{\mathbb{R}^+}2t\mathrm{d}t\cdot|g(\cdot)| \leq C\Big(\frac{1}{M\eta}\Big)^{C_2}\Big(\sum_{j=1}^W\frac{r_{j,1} + r_{j,2}}{1 + 2t_j^2}\Big)^{-C_2}\prod_{k=p,q}\frac{(1 + 2t_k^2)^{C_1}}{\big((\min\{\mathsf{Re}b_{k,1}, \mathsf{Re}b_{k,2}\}\big)^{C_1}}. \tag{6.36}$$

Now, we use the trivial bounds

$$\Big(\sum_{j=1}^W\frac{r_{j,1} + r_{j,2}}{1 + 2t_j^2}\Big)^{-C_2} \leq \Big(\frac{1 + 2t_p^2}{r_{p,1} + r_{p,2}}\Big)^{C_2} \leq \Big((1 + 2t_p^2)(r_{p,1}^{-1} + r_{p,2}^{-1})\Big)^{C_2}, \tag{6.37}$$

and

$$\frac{1 + 2t_k^2}{\min\{\mathsf{Re}b_{k,1}, \mathsf{Re}b_{k,2}\}} \leq C(1 + 2t_k^2)(r_{k,1}^{-1} + r_{k,2}^{-1}). \tag{6.38}$$

Inserting (6.37) and (6.38) into (6.36) and integrating out the remaining variables yields

$$|\mathbb{G}(\hat{B},T)| \le C\Big(\frac{1}{M\eta}\Big)^{C_1} \prod_{k=p,q} \Big(r_{k,1}^{-1}+r_{k,2}^{-1}+t_k+1\Big)^{C_3}. \tag{6.39}$$

Combining (6.25) and (6.39) we can get the bound

$$|\mathsf{F}(\hat{X},\hat{B},V,T)| \le e^{O(WN^{\varepsilon_2})} \prod_{k=p,q} \Big(r_{k,1}^{-1}+r_{k,2}^{-1}+t_k+1\Big)^{O(1)}.$$

Hence, we completed the proof of Lemma 6.5. □

6.4. **Summing up: Proof of Lemma 6.1.** In the discussions in Sections 6.2 and 6.3, we ignored the irrelevant factor $\mathsf{Q}(\cdot)$. However, it is easy to modify the discussion slightly to take this factor into account, whereby we can prove Lemma 6.1.

Proof of Lemma 6.1. At first, by the definition in (6.4), we can rewrite (3.31) as

$$\mathsf{A}(\cdot) = \int \mathrm{d}X^{[1]}\mathrm{d}\mathbf{y}^{[1]}\mathrm{d}\mathbf{w}^{[1]}\mathrm{d}\Omega\mathrm{d}\Xi\mathrm{d}\mu(P_1)\mathrm{d}\nu(Q_1)\ \mathcal{P}(\cdot)\mathsf{Q}(\cdot)\mathcal{F}(\cdot).$$

Now, by the conclusion $\kappa_1 = W^{O(1)}$ in Lemma 6.3, it suffices to consider one term in $\mathsf{Q}(\cdot)$, which is a monomial of the form

$$\mathfrak{p}(t,s,(y_p^{[1]})^{-1},(y_q^{[1]})^{-1})\mathfrak{q}(\Omega,\Xi),$$

regarding σ, $v_p^{[1]}$, $v_q^{[1]}$, P_1-variables, $X^{[1]}$-variables and $\mathbf{w}^{[1]}$-variables as bounded parameters. Here $\mathfrak{p}(\cdot)$ is a monomial of $t,s,(y_p^{[1]})^{-1},(y_q^{[1]})^{-1}$ and $\mathfrak{q}(\cdot)$ is a monomial of Ω,Ξ-variables, both with bounded coefficients and bounded degrees, according to the fact $\kappa_2,\kappa_3 = O(1)$ in Lemma 6.3. Now we define

$$\begin{aligned}
\mathsf{P}_{\mathfrak{q}}(\hat{X},\hat{B},V,T) &:= \int \mathrm{d}\Omega\mathrm{d}\Xi\ \mathcal{P}(\Omega,\Xi,\hat{X},\hat{B},V,T)\cdot\mathfrak{q}(\Omega,\Xi),\\
\mathsf{F}_{\mathfrak{p}}(\hat{X},\hat{B},V,T) &:= \int \mathrm{d}X^{[1]}\mathrm{d}\mathbf{y}^{[1]}\mathrm{d}\mathbf{w}^{[1]}\mathrm{d}\mu(P_1)\mathrm{d}\nu(Q_1)\\
&\quad\times\mathcal{F}(\hat{X},\hat{B},V,T,P_1,Q_1,X^{[1]},\mathbf{y}^{[1]},\mathbf{w}^{[1]})\cdot\mathfrak{p}(t,s,(y_p^{[1]})^{-1},(y_q^{[1]})^{-1}).
\end{aligned} \tag{6.40}$$

By repeating the discussions in Sections 6.2 and 6.3 with slight modification, we can easily see that

$$\begin{aligned}
|\mathsf{P}_{\mathfrak{q}}(\hat{X},\hat{B},V,T)| &\le e^{O(W)}\prod_{j=1}^{W}(r_{j,1}^{-1}+r_{j,2}^{-1}+t_j+1)^{O(1)},\\
|\mathsf{F}_{\mathfrak{p}}(\hat{X},\hat{B},V,T)| &\le e^{O(WN^{\varepsilon_2})}\prod_{k=p,q}\Big(r_{k,1}^{-1}+r_{k,2}^{-1}+t_k+1\Big)^{O(1)}
\end{aligned}$$

hold as well. Therefore, we completed the proof of Lemma 6.1. □

7. Proofs of Lemmas 5.1 and 5.6

In this section, with the aid of Lemma 6.1, we prove Lemmas 5.1 and 5.6. According to Lemmas 5.2 and 5.3, one can see that away from the saddles, $\mathsf{Re}\mathring{L}(\hat{B},T)$ and $\mathsf{Re}\mathring{K}(\hat{X},V)$ increase quadratically in $\hat{B}$-variables and $\hat{X}$-variables, respectively. Hence, it is easy to control the integral (5.2) over these variables outside the vicinities. However, like $\hat{B}$-variables, the domain of $\mathbf{t}$-variables is also not compact. This forces us to analyze the exponential function

$$\mathbb{M}(\mathbf{t}) := \exp\Big\{-M\mathsf{Re}\big(\ell_S(\hat{B},T)\big)\Big\} \tag{7.1}$$

carefully for any fixed $\hat{B}$-variables.

Recall the definition of the sector $\mathbb{K}$ in (6.2). For $\mathbf{b}_1 \in \mathbb{K}^W$ and $\mathbf{b}_2 \in \bar{\mathbb{K}}^W$, we have

$$\min_{j,k} \mathsf{Re}(b_{j,1}+b_{j,2})(b_{k,1}+b_{k,2}) \geq c \min_{j,k} \sum_{a,b=1,2} r_{j,a} r_{k,b} \geq c \min_{j,a} r_{j,a}^2 := \mathfrak{A}(\hat{B}), \tag{7.2}$$

for some positive constant c depending on κ from (1.15). From now on, we regard $\mathbb{M}(\mathbf{t})$ as a measure of the $\mathbf{t}$-variables and study it in the following two regions separately:

$$(i): \mathbf{t} \in \mathbb{I}^{W-1}, \qquad (ii): \mathbf{t} \in \mathbb{R}_+^{W-1} \setminus \mathbb{I}^{W-1}.$$

Roughly speaking, when $\mathbf{t} \in \mathbb{I}^{W-1}$, we will see that $\mathbb{M}(\mathbf{t})$ can be bounded pointwisely by a Gaussian measure. More specifically, we have the following lemma.

Lemma 7.1. *With the notation above, we have*

$$\mathbb{M}(\mathbf{t}) \leq \exp\Big\{ -\frac{M}{12}\mathfrak{A}(\hat{B}) \sum_{j,k} \mathfrak{s}_{jk}(t_k - t_j)^2 \Big\}, \quad \forall\, \mathbf{t} \in \mathbb{I}^{W-1}.$$

However, the behavior of $\mathbb{M}(\mathbf{t})$ for $\mathbf{t} \in \mathbb{R}_+^{W-1} \setminus \mathbb{I}^{W-1}$ is much more sophisticated. We will not try to provide a pointwise control of $\mathbb{M}(\mathbf{t})$ in this region. Instead, we will bound the integral of $\mathfrak{q}(\mathbf{t})$ against $\mathbb{M}(\mathbf{t})$ over this region, for any given monomial $\mathfrak{q}(\cdot)$ of interest. More specifically, recalling the definition of Θ in (5.33) and the spanning tree $\mathcal{G}_0 = (\mathcal{V}, \mathcal{E}_0)$ in Assumption 1.1, and additionally setting

$$\mathfrak{L} := \frac{M}{4}\mathfrak{A}(\hat{B}) \min_{i,j \in \mathcal{E}_0} \mathfrak{s}_{ij}, \tag{7.3}$$

we have the following lemma.

Lemma 7.2. *Let $\mathfrak{q}(\mathbf{t}) = \prod_{j=2}^W t_j^{n_j}$ be a monomial of $\mathbf{t}$-variables, with powers $n_j = O(1)$ for all $j = 2, \ldots, W$. We have*

$$\int_{\mathbb{R}_+^{W-1} \setminus \mathbb{I}^{W-1}} \prod_{j=2}^W \mathrm{d}t_j\, \mathbb{M}(\mathbf{t})\mathfrak{q}(\mathbf{t}) \leq \Big(1 + \mathfrak{L}^{-\frac{1}{2}}\Big)^{O(W^2)} \exp\Big\{ -\Theta^2 \mathfrak{A}(\hat{B}) + O(W^2 \log N) \Big\} \tag{7.4}$$

Remark 7.3. *Roughly speaking, by Lemma 7.2 we see that the integral of $\mathfrak{q}(\mathbf{t})$-variables against the measure $\mathbb{M}(\mathbf{t})$ over the region $\mathbb{R}_+^{W-1} \setminus \mathbb{I}^{W-1}$ is exponentially small, owing to the fact $\Theta^2 \gg W^2 \log N$.*

We will postpone the proofs of Lemmas 7.1 and 7.2 to the end of this section. In the sequel, at first, we prove Lemmas 5.1 and 5.6 with the aid of Lemmas 6.1, 7.1 and 7.2. Before commencing the formal proofs, we mention two basic facts which are formulated as the following lemma.

Lemma 7.4. *Under Assumption 1.1, we have the following two facts.*

- *For the smallest eigenvalue of $S^{(1)}$, there exists some positive constant c such that*

$$\lambda_1(-S^{(1)}) \geq \frac{c}{W^2}. \tag{7.5}$$

- *Let $\boldsymbol{\varrho} = (\rho_2, \ldots, \rho_W)'$ be a real vector and $\rho_1 = 0$. If there is at least one $\alpha \in \{2, \ldots, W\}$ such that $\varrho_\alpha \geq \Theta/\sqrt{M}$, then we have*

$$\sum_{j,k} \mathfrak{s}_{jk}(\varrho_j - \varrho_k)^2 \geq \frac{\Theta}{M}. \tag{7.6}$$

Proof. Let $\boldsymbol{\varrho} = (\rho_2, \ldots, \rho_W)'$ be a real vector and $\rho_1 = 0$. Now, we assume $|\rho_\alpha| = \max_{\beta=2,\ldots,W} |\rho_\beta|$. Then

$$\frac{-\boldsymbol{\rho}' S^{(1)} \boldsymbol{\rho}}{||\boldsymbol{\rho}||_2^2} = \frac{\frac{1}{2}\sum_{j,k} \mathfrak{s}_{jk}(\rho_j - \rho_k)^2}{\sum_j \rho_j^2} \geq \frac{c(\rho_\alpha - \rho_1)^2}{W^2 \rho_\alpha^2} = \frac{c}{W^2},$$

where the second step follows from Assumption 1.1 (iv) and Cauchy-Schwarz inequality. Analogously, we have

$$\sum_{j,k}\mathfrak{s}_{jk}(\varrho_j-\varrho_k)^2\geq\frac{c}{W}\varrho_\alpha^2\geq\frac{\Theta}{M}$$

according to the definition of Θ in (5.33). Hence, we completed the proof. □

Recalling the notation defined in (5.2) and the facts $|x_{j,a}|=1$ and $|b_{j,a}|=r_{j,a}$ for all $j=1,\ldots,W$ and $a=1,2$, for any sequence of domains, we have

$$\begin{aligned}&\big|\mathcal{I}(\mathbf{I}_1^b,\mathbf{I}_2^b,\mathbf{I}_1^x,\mathbf{I}_2^x,\mathbf{I}^t,\mathbf{I}^v)\big|\\&\leq e^{O(W\log N)}\int_{\mathbf{I}_1^b}\prod_{j=1}^W\mathrm{d}b_{j,1}\int_{\mathbf{I}_2^b}\prod_{j=1}^W\mathrm{d}b_{j,2}\int_{\mathbf{I}_1^x}\prod_{j=1}^W\mathrm{d}x_{j,1}\int_{\mathbf{I}_2^x}\prod_{j=1}^W\mathrm{d}x_{j,2}\int_{\mathbf{I}^t}\prod_{j=2}^W2t_j\mathrm{d}t_j\int_{\mathbf{I}^v}\prod_{j=2}^W2v_j\mathrm{d}v_j\\&\quad\times\exp\Big\{-M\big(\mathsf{Re}K(\hat{X},V)+\mathsf{Re}L(\hat{B},T)\big)\Big\}\cdot|\mathsf{A}(\hat{X},\hat{B},V,T)|\cdot\prod_{j=1}^W(r_{j,1}+r_{j,2})^2.\end{aligned}\tag{7.7}$$

In addition, according to Lemma 6.1, we have

$$|\mathsf{A}(\hat{X},\hat{B},V,T)|\cdot\prod_{j=2}^W2t_j\cdot\prod_{j=2}^W2v_j\cdot\prod_{j=1}^W(r_{j,1}+r_{j,2})^2\leq e^{O(WN^{\varepsilon_2})}\tilde{\mathfrak{p}}(\mathbf{r},\mathbf{r}^{-1},\mathbf{t}),\tag{7.8}$$

for some polynomial $\tilde{\mathfrak{p}}(\mathbf{r},\mathbf{r}^{-1},\mathbf{t})$ with positive coefficients, and

$$\tilde{\mathfrak{p}}(\mathbf{r},\mathbf{r}^{-1},\mathbf{t})\in\mathfrak{Q}\Big(\{r_{j,1},r_{j,2},r_{j,1}^{-1},r_{j,2}^{-1},t_j\}_{j=1}^W;\kappa_1,\kappa_2,\kappa_3\Big),\quad\kappa_1=e^{O(W)},\quad\kappa_2,\kappa_3=O(1).\tag{7.9}$$

7.1. **Proof of Lemma 5.1.** At first, since throughout the whole proof, the domains of $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{v}$-variables, namely, Σ^W, Σ^W and $\mathbb{I}^{W-1}$, will not be involved, we just use $*$'s to represent them, in order to simplify the notation.

Now, we introduce the following contours with the parameter $\mathfrak{D}\in\mathbb{R}_+$,

$$\Gamma_{\mathfrak{D}}:=\big\{ra_+|r\in[0,\mathfrak{D}]\big\}\subset\Gamma,\quad\mathbb{R}_{\mathfrak{D}}=[0,(\mathsf{Re}a_+)\mathfrak{D}]\subset\mathbb{R}_+,\quad\mathcal{L}_{\mathfrak{D}}:=\big\{(\mathsf{Re}a_+)\mathfrak{D}+\mathbf{i}(\mathsf{Im}a_+)r|r\in[0,\mathfrak{D}]\big\}.$$

In addition, we recall the sector $\mathbb{K}$ defined in (6.2). Then, trivially, we have

$$\mathbb{R}_+,\Gamma,\mathcal{L}_{\mathfrak{D}}\subset\mathbb{K},\quad\mathbb{R}_+,\bar{\Gamma},\bar{\mathcal{L}}_{\mathfrak{D}}\in\bar{\mathbb{K}},\quad\forall\,\mathfrak{D}\in\mathbb{R}_+.$$

We claim that the integrand in (5.2) is an analytic function of the $\hat{B}$-variables. To see this, we can go back to the integral representation (3.17) and the definitions of $L(B)$ and $\mathcal{P}(\Omega,\Xi,X,B)$ in (3.18). Note that since $\exp\{M\log\det B_j\}=(\det B_j)^M$, actually the logarithmic terms in $L(B)$ do not produce any singularity in the integrand in (3.17). In addition, according to the fact that the $\chi^\ell=0$ for any Grassmann variable χ and $\ell\geq2$, the factors $\det{}^M(1+M^{-1}X_j^{-1}\Omega_jB_j^{-1}\Xi_j)$ is actually a polynomial of Ω_j, Ξ_j, X_j^{-1} and B_j^{-1}-entries with degree 16. The other factors containing $\hat{B}$-variables $\mathcal{P}(\cdot)$ can be checked analogously. Hence, it is easy to see that $\exp\{-ML(\cdot)\}\mathcal{P}(\cdot)$ is analytic in $\hat{B}$-variables. Consequently, we have

$$\mathcal{I}\Big((\Gamma_{\mathfrak{D}}\cup\mathcal{L}_{\mathfrak{D}})^W,(\bar{\Gamma}_{\mathfrak{D}}\cup\bar{\mathcal{L}}_{\mathfrak{D}})^W,*,*,\mathbb{R}_+^{W-1},*\Big)=\mathcal{I}\Big((\mathbb{R}_{\mathfrak{D}})^W,(\mathbb{R}_{\mathfrak{D}})^W,*,*,\mathbb{R}_+^{W-1},*\Big).$$

Hence, to prove Lemma 5.1, it suffices to prove the following lemma.

Lemma 7.5. *Suppose that $|E|\leq\sqrt{2}-\kappa$. As $\mathfrak{D}\to\infty$, the following convergence hold,*

$$\begin{aligned}&(i):\quad\mathcal{I}\Big((\Gamma_{\mathfrak{D}}\cup\mathcal{L}_{\mathfrak{D}})^W,(\bar{\Gamma}_{\mathfrak{D}}\cup\bar{\mathcal{L}}_{\mathfrak{D}})^W,*,*,\mathbb{R}_+^{W-1},*\Big)-\mathcal{I}\Big((\Gamma_{\mathfrak{D}})^W,(\bar{\Gamma}_{\mathfrak{D}})^W,*,*,\mathbb{R}_+^{W-1},*\Big)\to0,\\&(ii):\quad\mathcal{I}\Big((\Gamma)^W,(\bar{\Gamma})^W,*,*,\mathbb{R}_+^{W-1},*\Big)-\mathcal{I}\Big((\Gamma_{\mathfrak{D}})^W,(\bar{\Gamma}_{\mathfrak{D}})^W,*,*,\mathbb{R}_+^{W-1},*\Big)\to0,\end{aligned}$$

$$(iii): \quad \mathcal{I}\Big(\mathbb{R}_+^W, \mathbb{R}_+^W, *, *, \mathbb{R}_+^{W-1}, *\Big) - \mathcal{I}\Big(\mathbb{R}_{\mathfrak{D}}^W, \mathbb{R}_{\mathfrak{D}}^W, *, *, \mathbb{R}_+^{W-1}, *\Big) \to 0.$$

Proof. For simplicity, we use the notation

$$\begin{aligned}
\mathbf{I}_{\mathfrak{D}}^{b,1} &:= \big(\Gamma_{\mathfrak{D}} \cup \mathcal{L}_{\mathfrak{D}}\big)^W \times \big(\bar{\Gamma}_{\mathfrak{D}} \cup \bar{\mathcal{L}}_{\mathfrak{D}}\big)^W \setminus (\Gamma_{\mathfrak{D}})^W \times (\bar{\Gamma}_{\mathfrak{D}})^W,\\
\mathbf{I}_{\mathfrak{D}}^{b,2} &:= \Gamma^W \times \bar{\Gamma}^W \setminus (\Gamma_{\mathfrak{D}})^W \times (\bar{\Gamma}_{\mathfrak{D}})^W,\\
\mathbf{I}_{\mathfrak{D}}^{b,3} &:= \mathbb{R}_+^W \times \mathbb{R}_+^W \setminus \mathbb{R}_{\mathfrak{D}}^W \times \mathbb{R}_{\mathfrak{D}}^W.
\end{aligned}$$

Now, recall the definition of the function $\ell(\mathbf{a})$ in (5.6) and the representation of $L(\hat{B}, T)$ in (5.5). Hence, in light of the definition of $\mathbb{M}(\mathbf{t})$ in (7.1), we have

$$\exp\big\{-M\mathsf{Re}L(\hat{B},T)\big\} = \exp\big\{-M\big(\mathsf{Re}\ell(\mathbf{b}_1) + \mathsf{Re}\ell(-\mathbf{b}_2)\big)\big\}\mathbb{M}(\mathbf{t}). \tag{7.10}$$

By the assumption $|E| \le \sqrt{2} - \kappa$, we see that $\mathsf{Re}b_{j,a}b_{k,a} > 0$ for all $b_{j,a}, b_{k,a} \in \mathbb{K} \cup \bar{\mathbb{K}}$. Consequently, when $b_{j,1} \in \mathbb{K}$ and $b_{j,2} \in \bar{\mathbb{K}}$ for all $j = 1, \ldots, W$, we have for some positive constant c dependent of κ in (1.15),

$$\begin{aligned}
\mathsf{Re}\ell(\mathbf{b}_1) + \mathsf{Re}\ell(-\mathbf{b}_2) &\ge \sum_{a=1,2}\sum_j \Big(\frac{1}{2}(1+\mathfrak{s}_{jj})\mathsf{Re}b_{j,a}^2 + (-1)^{a+1}E\mathsf{Im}b_{j,a} - \log r_{j,a}\Big)\\
&\ge c\sum_{a=1,2}\sum_j r_{j,a}^2 - \sum_{a=1,2}\sum_j \log r_{j,a},
\end{aligned} \tag{7.11}$$

where we used Assumption 1.1 (ii) and the fact that $(-1)^{a+1}E\mathsf{Im}b_{j,a} \ge 0$.

Now, when $(\mathbf{b}_1, \mathbf{b}_2) \in \mathbf{I}_{\mathfrak{D}}^{b,i}$ for $i = 1, 2, 3$, we have $\sum_{a=1,2}\sum_j r_{j,a}^2 \ge c\mathfrak{D}^2$ for some positive constant c, which implies the trivial fact

$$\sum_{a=1,2}\sum_j r_{j,a}^2 \ge \frac{1}{2}\sum_{a=1,2}\sum_j r_{j,a}^2 + \frac{c}{2}\mathfrak{D}^2. \tag{7.12}$$

Consequently, we can get from (7.10), (7.11) and (7.12) that for some positive constant c,

$$\exp\big\{-M\mathsf{Re}L(\hat{B},T)\big\} \le e^{-cM\mathfrak{D}^2}\prod_{a=1,2}\prod_{j=1}^W e^{-cMr_{j,a}^2}\, r_{j,a}^M \cdot \mathbb{M}(\mathbf{t})$$

holds in $\mathbf{I}_{\mathfrak{D}}^{b,i}$ for $i = 1, 2, 3$. In addition, by the boundedness of V and $\hat{X}$-variables, we can get the trivial bound $MK(\hat{X}, V) = O(N)$. Hence, from (7.7) and (7.8) we see that the quantities in Lemma 7.5 (i), (ii) and (iii) can be bounded by the following integral with $i = 1, 2, 3$, respectively,

$$\begin{aligned}
&e^{O(W\log N)}\int_{\mathbf{I}_{\mathfrak{D}}^{b,i}}\prod_{j=1}^W \mathrm{d}b_{j,1}\mathrm{d}b_{j,2}\int_{\Sigma^W}\prod_{j=1}^W \mathrm{d}x_{j,1}\int_{\Sigma^W}\prod_{j=1}^W \mathrm{d}x_{j,2}\int_{\mathbb{R}_+^{W-1}}\prod_{j=2}^W 2t_j\mathrm{d}t_j\int_{\mathbb{I}^{W-1}}\prod_{j=2}^W 2v_j\mathrm{d}v_j\\
&\times \exp\Big\{-M\big(\mathsf{Re}K(\hat{X},V) + \mathsf{Re}L(\hat{B},T)\big)\Big\}\cdot|\mathsf{A}(\hat{X},\hat{B},V,T)|\cdot\prod_{j=1}^W(r_{j,1}+r_{j,2})^2\\
&\le e^{O(N)}e^{-cM\mathfrak{D}^2}\int_{\mathbf{I}_{\mathfrak{D}}^{b,i}}\prod_{j=1}^W \mathrm{d}b_{j,1}\mathrm{d}b_{j,2}\int_{\mathbb{R}_+^{W-1}}\prod_{j=2}^W \mathrm{d}t_j\prod_{a=1,2}\prod_{j=1}^W e^{-cMr_{j,a}^2}\, r_{j,a}^M\cdot\mathbb{M}(\mathbf{t})\cdot\tilde{\mathfrak{p}}(\mathbf{r},\mathbf{r}^{-1},\mathbf{t}).
\end{aligned} \tag{7.13}$$

According to the facts $\kappa_1 = e^{O(W)}$ and $\kappa_2 = O(1)$ in (7.9), it is suffices to consider one monomial in $\tilde{\mathfrak{p}}(\mathbf{r}, \mathbf{r}^{-1}, \mathbf{t})$ with bounded coefficient. That means, it suffices to estimate the integral

$$\int_{\mathbf{I}_{\mathfrak{D}}^{b,i}}\prod_{j=1}^W \mathrm{d}b_{j,1}\mathrm{d}b_{j,2}\int_{\mathbb{R}_+^{W-1}}\mathrm{d}t_j\prod_{a=1,2}\prod_{j=1}^W e^{-cMr_{j,a}^2}\, r_{j,a}^M\cdot\mathbb{M}(\mathbf{t})\cdot\tilde{\mathfrak{q}}(\mathbf{r},\mathbf{r}^{-1},\mathbf{t}), \quad i = 1,2,3, \tag{7.14}$$

for some monomial

$$\tilde{\mathfrak{q}}(\mathbf{r},\mathbf{r}^{-1},\mathbf{t}) = \prod_{a=1,2}\prod_{j=1}^W r_{j,a}^{\ell_{j,a}}\prod_{j=2}^W t_j^{n_j}, \quad |\ell_{j,a}|, n_j = O(1), \quad \text{for} \quad j = 1,\ldots,W, \quad a = 1,2, \tag{7.15}$$

where the bound on ℓ_j's and n_j's follows from the fact that $\kappa_3 = O(1)$ in (7.9).

Bounding t_j's by 1 trivially in the region $\mathbf{t} \in \mathbb{I}^{W-1}$ and using Lemma 7.2 in the region $\mathbf{t} \in \mathbb{R}_+^{W-1} \setminus \mathbb{I}^{W-1}$, we can get for $i = 1, 2, 3$,

$$(7.14) \le e^{O(W^2 \log N)} \int_{\mathbf{I}_{\mathfrak{D}}^{b,i}} \prod_{j=1}^{W} \mathrm{d}b_{j,1} \mathrm{d}b_{j,2} \cdot \prod_{a=1,2} \prod_{j=1}^{W} e^{-cMr_{j,a}^2} r_{j,a}^{M+\ell_{j,a}} \big(1 + \mathfrak{L}^{-\frac{1}{2}}\big)^{O(W^2)}.$$

By the definition of $\mathfrak{A}(\hat{B})$ in (7.2) and the assumption $M \gg W^4$ we see that

$$\prod_{a=1,2} \prod_{j=1}^{W} r_{j,a}^{M+\ell_{j,a}} \big(1 + \mathfrak{L}^{-\frac{1}{2}}\big)^{O(W^2)} \le \prod_{a=1,2} \prod_{j=1}^{W} r_{j,a}^{M(1+o(1))}.$$

Consequently, by using elementary Gaussian integral, we can get the trivial bound

$$(7.14) \le e^{O(N \log N)}, \quad i = 1, 2, 3,$$

and then we have for $i = 1, 2, 3$,

$$(7.13) \le e^{-cM\mathfrak{D}^2 + O(N \log N)} \to 0, \quad \text{as} \quad \mathfrak{D} \to \infty.$$

Thus we completed the proof. □

7.2. **Proof of Lemma 5.6.** Plugging the first identity of (5.20) and (7.8) into (7.7), we can write

$$\begin{aligned} \big|\mathcal{I}(\mathbf{I}_1^b, \mathbf{I}_2^b, \mathbf{I}_1^x, \mathbf{I}_2^x, \mathbf{I}^t, \mathbf{I}^v)\big| \le e^{O(WN^{\varepsilon_2})} & \int_{\mathbf{I}_1^b} \prod_{j=1}^{W} \mathrm{d}b_{j,1} \int_{\mathbf{I}_2^b} \prod_{j=1}^{W} \mathrm{d}b_{j,2} \int_{\mathbf{I}_1^x} \prod_{j=1}^{W} \mathrm{d}x_{j,1} \int_{\mathbf{I}_2^x} \prod_{j=1}^{W} \mathrm{d}x_{j,2} \\ \times \int_{\mathbf{I}^t} \prod_{j=2}^{W} \mathrm{d}t_j & \int_{\mathbf{I}^v} \prod_{j=2}^{W} \mathrm{d}v_j \exp\Big\{-M\Big(\mathring{\mathsf{Re}K}(\hat{X}, V) + \mathring{\mathsf{Re}L}(\hat{B}, T)\Big)\Big\} \tilde{\mathfrak{p}}(\mathbf{r}, \mathbf{r}^{-1}, \mathbf{t}), \end{aligned} \tag{7.16}$$

where $\tilde{\mathfrak{p}}(\mathbf{r}, \mathbf{r}^{-1}, \mathbf{t})$ is specified in (7.9).

Lemma 5.6 immediately follows from the following two lemmas.

Lemma 7.6. *Under Assumptions 1.1 and 1.13, we have*

$$\mathcal{I}\big(\Gamma^W, \bar{\Gamma}^W, \Sigma^W, \Sigma^W, \mathbb{R}_+^{W-1} \setminus \mathbb{I}^{W-1}, \mathbb{I}^{W-1}\big) \le e^{-\Theta^2}. \tag{7.17}$$

Lemma 7.7. *Under Assumptions 1.1 and 1.13, we have*

$$\begin{aligned} \mathcal{I}\big(\Gamma^W, \bar{\Gamma}^W, \Sigma^W, \Sigma^W, \mathbb{I}^{W-1}, \mathbb{I}^{W-1}\big) = 2^W & \mathcal{I}\big(\Upsilon_+^b, \Upsilon_-^b, \Upsilon_+^x, \Upsilon_-^x, \Upsilon_S, \Upsilon_S\big) \\ & + \mathcal{I}\big(\Upsilon_+^b, \Upsilon_-^b, \Upsilon_+^x, \Upsilon_+^x, \Upsilon_S, \mathbb{I}^{W-1}\big) \\ & + \mathcal{I}\big(\Upsilon_+^b, \Upsilon_-^b, \Upsilon_-^x, \Upsilon_-^x, \Upsilon_S, \mathbb{I}^{W-1}\big) + O(e^{-\Theta}). \end{aligned} \tag{7.18}$$

In the sequel, we prove Lemmas 7.6 and 7.7.

Proof of Lemma 7.6. Recall (7.16) with the choice of the integration domains

$$\big(\mathbf{I}_1^b, \mathbf{I}_2^b, \mathbf{I}_1^x, \mathbf{I}_2^x, \mathbf{I}^t, \mathbf{I}^v\big) = \big(\Gamma^W, \bar{\Gamma}^W, \Sigma^W, \Sigma^W, \mathbb{R}_+^{W-1} \setminus \mathbb{I}^{W-1}, \mathbb{I}^{W-1}\big).$$

To simplify the integral on the r.h.s. of (7.16), we use the fact $\mathring{\mathsf{Re}K}(\hat{X}, V) \ge 0$ implied by (5.26), together with the facts that the $\mathbf{x}$ and $\mathbf{v}$-variables are bounded by 1. Consequently, we can eliminate the integral over $\mathbf{x}$ and $\mathbf{v}$-variables from the integral on the r.h.s. of (7.16). Moreover, according to (7.9), it suffices to prove

$$\int_{\Gamma^W} \prod_{j=1}^{W} \mathrm{d}b_{j,1} \int_{\bar{\Gamma}^W} \prod_{j=1}^{W} \mathrm{d}b_{j,2} \int_{\mathbb{R}_+^{W-1} \setminus \mathbb{I}^{W-1}} \prod_{j=2}^{W} \mathrm{d}t_j \exp\Big\{-M\mathring{\mathsf{Re}L}(\hat{B}, T)\Big\} \cdot \tilde{\mathfrak{q}}(\mathbf{r}, \mathbf{r}^{-1}, \mathbf{t}) \le e^{-\Theta^2} \tag{7.19}$$

instead, where $\tilde{\mathfrak{q}}(\cdot)$ is the monomial defined in (7.15).

Now, by the first inequality of (5.13), we have

$$\exp\big\{-M\mathring{\mathsf{Re}L}(\hat{B},T)\big\}\le\exp\Big\{-M\sum_{a=1,2}\sum_{j=1}^{W}\Big(c(r_{j,a}-1)^2+(r_{j,a}-\log r_{j,a}-1)\Big)\Big\}\cdot\mathbb{M}(\mathbf{t}). \tag{7.20}$$

At first, we integrate $\mathbf{t}$-variables out by using Lemma 7.2, namely,

$$\begin{aligned}&\int_{\mathbb{R}_+^{W-1}\setminus\mathbb{I}^{W-1}}\prod_{j=2}^{W}\mathrm{d}t_j\ \mathbb{M}(\mathbf{t})\cdot\tilde{\mathfrak{q}}(\mathbf{r},\mathbf{r}^{-1},\mathbf{t})\\&\le\prod_{a=1,2}\prod_{j=1}^{W}r_{j,a}^{\ell_{j,a}}\cdot\Big(1+\mathfrak{L}^{-\frac{1}{2}}\Big)^{O(W^2)}\exp\Big\{-\Theta^2\mathfrak{A}(\hat{B})+O(W^2\log N)\Big\}.\end{aligned} \tag{7.21}$$

Recall the definitions of $\mathfrak{L}$ and $\mathfrak{A}(\hat{B})$ in (7.3) and (7.2), and the fact $W^2=o(M^{1/2})$. The proof of the following fact is an elementary exercise

$$\sup_{r\in\mathbb{R}_+}\exp\big\{-M(r-\log r-1)\big\}r^{\ell}=O(1),\quad\text{if}\quad\ell=O(M^{\frac{1}{2}}). \tag{7.22}$$

Hence, we get the bound

$$\begin{aligned}&\prod_{a=1,2}\prod_{j=1}^{W}\exp\Big\{-M\big(r_{j,a}-\log r_{j,a}-1\big)\Big\}\cdot r_{j,a}^{\ell_{j,a}}\cdot\Big(1+\mathfrak{L}^{-\frac{1}{2}}\Big)^{O(W^2)}\\&\le\max_{b=1,2}\max_{k=1,\dots,W}\prod_{a=1,2}\prod_{j=1}^{W}\exp\Big\{-M\big(r_{j,a}-\log r_{j,a}-1\big)\Big\}\cdot r_{j,a}^{\ell_{j,a}}\Big(1+r_{k,b}^{-1}\Big)^{O(W^2)}=e^{O(W^2)}.\end{aligned} \tag{7.23}$$

Consequently, (7.20)-(7.23) imply that

$$\begin{aligned}&\int_{\mathbb{R}_+^{W-1}\setminus\mathbb{I}^{W-1}}\prod_{j=2}^{W}\mathrm{d}t_j\ \exp\Big\{-M\mathring{\mathsf{Re}L}(\hat{B},T)\Big\}\cdot\tilde{\mathfrak{q}}(\mathbf{r},\mathbf{r}^{-1},\mathbf{t})\\&\le\exp\Big\{-\Theta^2\mathfrak{A}(\hat{B})+O(W^2\log N)\Big\}\cdot\prod_{a=1,2}\prod_{j=1}^{W}\exp\big\{-cM(r_{j,a}-1)^2\big\}\\&\le e^{-\Theta^2}\cdot\prod_{a=1,2}\prod_{j=1}^{W}\exp\Big\{-\frac{c}{2}M(r_{j,a}-1)^2\Big\},\end{aligned} \tag{7.24}$$

for some positive constant c, where in the last step we use the obvious fact

$$\Theta^2\mathfrak{A}(\hat{B})+\frac{c}{2}\sum_{a=1,2}\sum_{j=1}^{W}M(r_{j,a}-1)^2\ge\Theta^2\gg W^2\log N$$

by (5.38) and the definition of $\mathfrak{A}(\hat{B})$ in (7.2). Plugging the bound (7.24) into the l.h.s of (7.19) and taking the integral over $\hat{B}$-variables we can see that (7.19) holds, which further implies (7.17). Therefore, we completed the proof of Lemma 7.6. □

To prove Lemma 7.7, we split the exponential function into two parts. We use one part to control the integral, and the other will be estimated by its magnitude. More specifically, we shall prove the following two lemmas.

Lemma 7.8. *Under Assumptions 1.1 and 1.13, we have*

$$\begin{aligned}&\int_{\Gamma^W}\prod_{j=1}^{W}\mathrm{d}b_{j,1}\int_{\bar{\Gamma}^W}\prod_{j=1}^{W}\mathrm{d}b_{j,2}\int_{\Sigma^W}\prod_{j=1}^{W}\mathrm{d}x_{j,1}\int_{\Sigma^W}\prod_{j=1}^{W}\mathrm{d}x_{j,2}\int_{\mathbb{I}^{W-1}}\prod_{j=2}^{W}\mathrm{d}t_j\int_{\mathbb{I}^{W-1}}\prod_{j=2}^{W}\mathrm{d}v_j\\&\times\exp\Big\{-\frac{1}{2}M\big(\mathsf{Re}\mathring{K}(\hat{X},V)+\mathsf{Re}\mathring{L}(\hat{B},T)\big)\Big\}\cdot\tilde{\mathfrak{p}}(\mathbf{r},\mathbf{r}^{-1},\mathbf{t})\le e^{O(W)}.\end{aligned} \tag{7.25}$$

Lemma 7.9. *If* $(\mathbf{b}_1, \mathbf{b}_2, \mathbf{x}_1, \mathbf{x}_2, \mathbf{t}, \mathbf{v}) \in \Gamma^W \times \bar{\Gamma}^W \times \Sigma^W \times \Sigma^W \times \mathbb{I}^{W-1} \times \mathbb{I}^{W-1}$, *but not in any of the Types I, II, III vicinities in Definition 5.5, we have*

$$\exp\Big\{ -\frac{1}{2} M \big(\mathsf{Re}\mathring{K}(\hat{X}, V) + \mathsf{Re}\mathring{L}(\hat{B}, T) \big) \Big\} \leq e^{-\Theta}. \tag{7.26}$$

With Lemmas 7.8 and 7.9, we can prove Lemma 7.7.

Proof of Lemma 7.7. For the sake of simplicity, in this proof, we temporarily use $\mathcal{I}_{\text{full}}$ to represent the l.h.s. of (7.18), i.e. the integral over the full domain, and use $\mathcal{I}_I$, $\mathcal{I}_{II}$ and $\mathcal{I}_{III}$ to represent the first three terms on the r.h.s. of (7.18). Now, combining (7.16), (7.26) and (7.25), we see that,

$$|\mathcal{I}_{\text{full}} - \mathcal{I}_I - \mathcal{I}_{II} - \mathcal{I}_{III}| \leq e^{O(WN^{\varepsilon_2})} \cdot e^{-\Theta} \cdot e^{O(W)} \leq e^{-\Theta},$$

in light of the definition of Θ in (5.33) and the assumption (5.37). Hence, we completed the proof of Lemma 7.7. □

Proof of Lemma 7.8. At first, again, the polynomial $\tilde{\mathfrak{p}}(\cdot)$ in the integrand can be replaced by the monomial $\tilde{\mathfrak{q}}(\cdot)$ defined in (7.15) in the discussion below, owing to the fact that $\kappa_1 = \exp\{O(W)\}$ in (7.9). Then, the proof is similar to that of Lemma 7.6, but much simpler, since $\mathbf{t}$-variables are bounded by 1 now. Consequently, we can eliminate $\hat{X}$, $\mathbf{t}$ and $\mathbf{v}$-variables from the integral directly and use the trivial bounds

$$\tilde{\mathfrak{q}}(\mathbf{r}, \mathbf{r}^{-1}, \mathbf{t}) \leq \prod_{a=1,2} \prod_{j=1}^{W} r_{j,a}^{\ell_{j,a}}, \qquad \mathsf{Re}\mathring{L}(\hat{B}, T) \;\geq\; c \sum_{a=1,2} \sum_{j=1}^{W} (r_{j,a} - 1)^2, \tag{7.27}$$

where the latter is from (5.13). Hence, it suffices to show

$$\int_{\Gamma^W} \prod_{j=1}^{W} \mathrm{d}b_{j,1} \int_{\bar{\Gamma}^W} \prod_{j=1}^{W} \mathrm{d}b_{j,2} \prod_{a=1,2} \prod_{j=1}^{W} \exp\left\{ -cM(r_{j,a} - 1)^2 \right\} r_{j,a}^{\ell_{j,a}} \leq e^{O(W)}. \tag{7.28}$$

Note that (7.28) follows from elementary Gaussian integral immediately. Therefore, we completed the proof of Lemma 7.8. □

Proof of Lemma 7.9. At first, according to (5.13) and (5.26), we see both $M\mathsf{Re}\mathring{L}(\hat{B}, T)$ and $M\mathsf{Re}\mathring{K}(\hat{X}, V)$ are nonnegative on the full domain. Hence, it suffices to show one of them is larger than Θ outside the Type I, II, III vicinities.

Note that for each type of vicinity, we have

$$(\mathbf{b}_1, \mathbf{b}_2, \mathbf{t}) \in \Upsilon_+^b \times \Upsilon_-^b \times \Upsilon_S. \tag{7.29}$$

Now, if (7.29) is violated, we have $(\mathbf{b}_1, \mathbf{b}_2) \in \Gamma^W \times \bar{\Gamma}^W \setminus \Upsilon_+^b \times \Upsilon_-^b$ or $\mathbf{t} \in \mathbb{I}^{W-1} \setminus \Upsilon_S$. If the former holds, by using (5.13) and the definition of Υ_+^b and Υ_-^b in (5.35), we have

$$M\mathsf{Re}\mathring{L}(\hat{B}, T) \geq cM \sum_{a=1,2} \sum_{j=1}^{W} (r_{j,a} - 1)^2 = cM||\mathbf{b}_1 - a_+||_2^2 + cM||\mathbf{b}_2 + a_-||_2^2 \geq \Theta,$$

which shows (7.26) if $(\mathbf{b}_1, \mathbf{b}_2) \in \Gamma^W \times \bar{\Gamma}^W \setminus \Upsilon_+^b \times \Upsilon_-^b$.

Hence, it suffices to consider the case $(\mathbf{b}_1, \mathbf{b}_2) \in \Upsilon_+^b \times \Upsilon_-^b$, $\mathbf{t} \in \mathbb{I}^{W-1} \setminus \Upsilon_S$. Using Lemma 7.1, we can see that

$$\mathbb{M}(\mathbf{t}) \leq \exp\left\{ -\frac{M}{6} \mathfrak{A}(\hat{B})(-\mathbf{t} S^{(1)} \mathbf{t}) \right\} \leq \exp\left\{ -\mathfrak{A}(\hat{B})\Theta \right\} \leq e^{-\Theta}, \tag{7.30}$$

where in the second step we used the definition of Υ_S in (5.35) and in the last step we used the fact $\mathfrak{A}(\hat{B}) \geq c$ if $(\mathbf{b}_1, \mathbf{b}_2) \in \Upsilon_+^b \times \Upsilon_-^b$. Then (7.20) and (7.30) also imply (7.26). Now, we turn to show

$M\mathsf{Re}\mathring{K}(\hat{X},V)\geq\Theta$ outside the vicinities. Recalling the definition of ϑ_j's in (5.25), we split the discussion into two cases

$$(i):\Big(\sin\vartheta_j-\frac{E}{2}\Big)^2\leq\frac{\Theta}{M},\quad\forall\, j=1,\dots,2W,\qquad (ii):\Big(\sin\vartheta_j-\frac{E}{2}\Big)^2>\frac{\Theta}{M},\quad\text{for some}\quad j\in\{1,\dots,2W\}.$$

Using (5.26), we can get $M\mathsf{Re}\mathring{K}(\hat{X},V)\geq\Theta$ in case (ii) immediately, so we can assume case (i) below. Then (i) implies that

$$|\arg(a_+^{-1}x_{j,a})|^2\wedge|\arg(a_-^{-1}x_{j,a})|^2\leq\frac{\Theta}{M},\quad\forall\, j=1,\dots,W;a=1,2. \tag{7.31}$$

Now, we claim that it suffices to focus on the following three subcases of (7.31),

(i') There exists a sequence of permutations of $\{1,2\}$, namely, $\boldsymbol{\epsilon}=(\epsilon_1,\dots,\epsilon_W)$, such that

$$|\arg(a_+^{-1}x_{j,\epsilon_j(1)})|^2\leq\frac{\Theta}{M},\quad|\arg(a_-^{-1}x_{j,\epsilon_j(2)})|^2\leq\frac{\Theta}{M},\quad\forall\, j=1,\dots,W.$$

(ii') There exists

$$|\arg(a_+^{-1}x_{j,a})|^2\leq\frac{\Theta}{M},\quad\forall\, j=1,\dots,W;a=1,2.$$

(iii') There exists

$$|\arg(a_-^{-1}x_{j,a})|^2\leq\frac{\Theta}{M},\quad\forall\, j=1,\dots,W;a=1,2.$$

To see this, note that for those $\hat{X}$-variables which satisfy (7.31) but do not belong to any of the case (i'), (ii') or (iii') listed above, there must be a pair $\{i,j\}\in\mathcal{E}$ such that

$$|\arg(a_+^{-1}x_{i,1})|^2,|\arg(a_+^{-1}x_{i,2})|^2\leq\frac{\Theta}{M},\quad|\arg(a_-^{-1}x_{j,1})|^2,|\arg(a_-^{-1}x_{j,2})|^2\leq\frac{\Theta}{M}, \tag{7.32}$$

or there exists a permutation ϵ_i such that

$$|\arg(a_+^{-1}x_{i,\epsilon_i(1)})|^2,|\arg(a_-^{-1}x_{i,\epsilon_i(2)})|^2\leq\frac{\Theta}{M},\quad|\arg(a_+^{-1}x_{j,1})|^2,|\arg(a_+^{-1}x_{j,2})|^2\leq\frac{\Theta}{M}, \tag{7.33}$$

or

$$|\arg(a_+^{-1}x_{i,\epsilon_i(1)})|^2,|\arg(a_-^{-1}x_{i,\epsilon_i(2)})|^2\leq\frac{\Theta}{M},\quad|\arg(a_-^{-1}x_{j,1})|^2,|\arg(a_-^{-1}x_{j,2})|^2\leq\frac{\Theta}{M}. \tag{7.34}$$

For each of (7.32), (7.33) and (7.34), we can perform a discussion similar to (5.30)-(5.31), to show that $M\mathsf{Re}\mathring{K}(\hat{X},V)\geq cM\gg\Theta$.

Hence, it suffices to focus on cases (i'), (ii') and (iii') in the sequel. Now, we denote the domains of $\hat{X}$-variables satisfying (i'), (ii') and (iii') by Υ_I^x, Υ_{II}^x and Υ_{III}^x, respectively. In addition, in the remaining part of this proof, we temporarily use the term Type A vicinity to represent its restriction on $\hat{X}$-variables, for $A=I,II,III$. Obviously, we have

$$\text{Type A vicinity}\subset\Upsilon_A^x,\qquad A=I,II,III.$$

If $(\mathbf{x}_1,\mathbf{x}_2)\in\Upsilon_I^x$ but outside the Type I vicinity,

$$||\arg(a_+^{-1}\mathbf{x}_{\boldsymbol{\epsilon}(1)})||_2^2+||\arg(a_-^{-1}\mathbf{x}_{\boldsymbol{\epsilon}(2)})||_2^2\geq\frac{\Theta}{M},\quad\forall\,\boldsymbol{\epsilon},$$

which easily implies that

$$M\mathsf{Re}\mathring{K}(\hat{X},V)\geq cM\sum_{j=1}^{2W}\Big(\sin\vartheta_j-\frac{E}{2}\Big)^2\geq\Theta. \tag{7.35}$$

Now, we turn to the case that there exists one sequence of permutations $\boldsymbol{\epsilon}$ such that $(\mathbf{x}_{\boldsymbol{\epsilon}(1)},\mathbf{x}_{\boldsymbol{\epsilon}(2)})\in\Upsilon_+^x\times\Upsilon_-^x$, but $\mathbf{v}_{\boldsymbol{\epsilon}}\notin\Upsilon_S$. In this case, we just go back to the first line of (5.15) and do the transform

$$V_j\to\mathfrak{J}V_j,\quad\hat{X}_j\to\mathfrak{J}\hat{X}_j\mathfrak{J}$$

for those j with $\epsilon(j) \neq \epsilon(1)$, where $\mathfrak{I}$ is defined in (4.1). Then, it suffices to consider

$$(\mathbf{x}_1, \mathbf{x}_2) \in \Upsilon_+^x \times \Upsilon_-^x, \quad \text{but} \quad \mathbf{v} \notin \Upsilon_S, \qquad \text{or} \qquad (\mathbf{x}_1, \mathbf{x}_2) \in \Upsilon_-^x \times \Upsilon_+^x, \quad \text{but} \quad \mathbf{v} \notin \Upsilon_S.$$

In either case, we can show that $M\mathsf{Re}\mathring{K}(\hat{X}, V) \geq \Theta$, analogously to case of $(\mathbf{b}_1, \mathbf{b}_2) \in \Upsilon_+^x \times \Upsilon_-^x$ but $\mathbf{t} \notin \Upsilon_S$, in (7.30).

Now, what remains is to show that for those $(\mathbf{x}_1, \mathbf{x}_2) \in \Upsilon_A^x$ but outside the Type A vicinity (A=II, III), we have $M\mathsf{Re}\mathring{K}(\hat{X}, V) \geq \Theta$. We only discuss the case $A = II$, the other is analogous. Note that outside the Type II vicinity of $\hat{X}$ variables we have

$$||\arg(a_+^{-1}\mathbf{x}_1)||_2^2 + ||\arg(a_+^{-1}\mathbf{x}_2)||_2^2 \geq \frac{\Theta}{M}. \tag{7.36}$$

Observe that now we are already in Υ_{II}^x, which means that all $x_{j,a}$'s are close to a_+ and far away from a_-. That means, we have $\sin(\arg(x_{ja})) - E/2 \sim \arg(a_+^{-1}x_{ja})$. Consequently, (7.36) also implies (7.35). Therefore, we completed the proof of Lemma 7.9. □

7.3. **Proof of Lemma 7.1.** Using the definition of $\ell_S(\hat{B}, T)$ in (5.6) and $\mathfrak{A}(\hat{B})$ in (7.2) and the fact $|(T_k T_j^{-1})_{12}| = |s_j t_k e^{\mathbf{i}\sigma_k} - s_k t_j e^{\mathbf{i}\sigma_j}|$, we have

$$\mathsf{Re}\ell_S(\hat{B}, T) \geq \frac{1}{2}\mathfrak{A}(\hat{B}) \sum_{j,k} \mathfrak{s}_{jk} |s_j t_k e^{\mathbf{i}\sigma_k} - s_k t_j e^{\mathbf{i}\sigma_j}|^2. \tag{7.37}$$

Simple estimate using $s_j^2 = 1 + t_j^2$ shows that

$$|s_j t_k e^{\mathbf{i}\sigma_k} - s_k t_j e^{\mathbf{i}\sigma_j}|^2 \geq \frac{1}{4}(t_k - t_j)^2 \Big(\frac{1}{1 + 2t_j^2} + \frac{1}{1 + 2t_k^2} \Big) \geq \frac{1}{6}(t_k - t_j)^2. \tag{7.38}$$

Notice that the assumption $\mathbf{t} \in \mathbb{I}^{W-1}$ was used only in the last inequality. By (7.37), (7.38) and the definition (7.1), Lemma 7.1 follows immediately.

7.4. **Proof of Lemma 7.2.** Let $\mathbb{I}^c = \mathbb{R}_+ \setminus \mathbb{I}$. Now we consider the domain sequence $\vec{\mathbb{J}} = (\mathbb{J}_2, \ldots, \mathbb{J}_W) \in \{\mathbb{I}, \mathbb{I}^c\}^{W-1}$. We decompose the integral in Lemma 7.2 as follows

$$\int_{\mathbb{R}^{W-1} \setminus \mathbb{I}^{W-1}} \prod_{j=2}^{W} \mathrm{d}t_j \, \mathbb{M}(\mathbf{t}) \mathfrak{q}(\mathbf{t}) = \sum_{\substack{\vec{\mathbb{J}} \in \{\mathbb{I}, \mathbb{I}^c\}^{W-1} \\ \vec{\mathbb{J}} \neq \mathbb{I}^{W-1}}} \int_{\prod_{j=2}^{W} \mathbb{J}_j} \prod_{j=2}^{W} \mathrm{d}t_j \, \mathbb{M}(\mathbf{t}) \mathfrak{q}(\mathbf{t}). \tag{7.39}$$

Note the total number of the choices of such $\vec{\mathbb{J}}$ in the sum above is $2^{W-1} - 1$. It suffices to consider one of these sequences $\vec{\mathbb{J}} \in \{\mathbb{I}, \mathbb{I}^c\}^{W-1}$ in which there is at least one i such that $\mathbb{J}_i = \mathbb{I}^c$.

Recall the spanning tree $\mathcal{G}_0 = (\mathcal{V}, \mathcal{E}_0)$ in Assumption 1.1. The simplest case is that there exists a linear spanning tree (a path) $\mathcal{G}_0$ with

$$\mathcal{E}_0 = \{(i, i+1)\}_{i=1}^{W-1} \subset \mathcal{E}. \tag{7.40}$$

We first present the proof in this simplest case.

Now, we only keep the edges in the path $\mathcal{E}_0$, i.e. the terms with $k = j - 1$ in (7.37), we also trivially discard the term $1/(1 + 2t_j^2)$ from the sum $1/(1 + 2t_{j-1}^2) + 1/(1 + 2t_j^2)$ in the estimate (7.38) (the first inequality), and finally we bound all $M\mathfrak{A}(\hat{B})\mathfrak{s}_{j-1,j}/4$ by $\mathfrak{L}$ defined in (7.3) from below. That means, we use the bound

$$\mathbb{M}(\mathbf{t}) \leq \prod_{j=2}^{W} \exp\Big\{ -\mathfrak{L}\frac{(t_j - t_{j-1})^2}{1 + 2t_{j-1}^2} \Big\} := \prod_{j=2}^{W} \breve{\mathbb{M}}_j(\mathbf{t}). \tag{7.41}$$

Consequently, we have

$$\int_{\prod_{j=2}^W \mathbb{J}_j} \prod_{j=2}^W \mathrm{d}t_j \, \mathbb{M}(\mathbf{t})\mathfrak{q}(\mathbf{t}) \le \int_{\prod_{j=2}^W \mathbb{J}_j} \prod_{j=2}^W \mathrm{d}t_j \, \prod_{j=2}^W t_j^{n_j} \breve{\mathbb{M}}_j(\mathbf{t}). \tag{7.42}$$

Note that, as a function of $\mathbf{t}$, $\breve{\mathbb{M}}_j(\mathbf{t})$ only depends on t_{j-1} and t_j.

Having fixed $\vec{\mathbb{J}}$, assume that k is the largest index such that $\mathbb{J}_k = \mathbb{I}^c$, i.e. $t_{k+1}, \dots, t_W \in \mathbb{I}$. Now, we claim that

$$\sum_{j=2}^W \frac{(t_j - t_{j-1})^2}{1 + 2t_{j-1}^2} \ge \sum_{j=2}^k \frac{(t_j - t_{j-1})^2}{1 + 2t_{j-1}^2} \ge \frac{1}{300k^2}, \quad \text{if} \quad t_k \in \mathbb{I}^c. \tag{7.43}$$

To see (7.43), we use the following elementary facts

$$\frac{(t_j - t_{j-1})^2}{1 + 2t_{j-1}^2} \ge \frac{1}{3}\frac{(t_j - t_{j-1})^2}{t_{j-1}^2} = \frac{1}{3}(t_j/t_{j-1} - 1)^2, \quad \text{if} \quad t_{j-1} \in \mathbb{I}^c \tag{7.44}$$

and

$$\frac{(t_j - t_{j-1})^2}{1 + 2t_{j-1}^2} \ge \frac{1}{3}(t_j - t_{j-1})^2, \quad \text{if} \quad t_{j-1} \in \mathbb{I} \tag{7.45}$$

for all $j = 2, \dots, W$. We show (7.43) by contradiction. If (7.43) is violated, we have

$$\frac{(t_j - t_{j-1})^2}{1 + 2t_{j-1}^2} \le \frac{1}{300k^2}, \quad \forall\, j = 2, \dots, k,$$

which together with (7.44) and (7.45) implies that

$$t_j \le t_{j-1}\Big(1 + \frac{1}{10k}\Big) + \frac{1}{10k}. \tag{7.46}$$

Using (7.46) recursively yields

$$t_k \le \Big(1 + \frac{1}{10k}\Big)^{k-1}(t_1 + 1) - 1 = \Big(1 + \frac{1}{10k}\Big)^{k-1} - 1 \le \frac{1}{2}, \tag{7.47}$$

where in the second step we used the fact $t_1 = 0$. Note that (7.47) contradicts $t_k \in \mathbb{I}^c$. Hence, we verified (7.43).

Now, we split $\prod_{j=2}^W \breve{\mathbb{M}}_j(\mathbf{t})$ into two parts. We use one to control the integral, and the other will be estimated by (7.43). Specifically, substituting (7.43) into (7.42) we have

$$\int_{\prod_{j=2}^W \mathbb{J}_j} \prod_{j=2}^W \mathrm{d}t_j \, \mathbb{M}(\mathbf{t})\mathfrak{q}(\mathbf{t}) \le e^{-\frac{\mathfrak{L}}{600k^2}} \int_{\mathbb{R}_+^{W-1}} \prod_{j=2}^W \mathrm{d}t_j \, \prod_{j=2}^W t_j^{n_j} \big(\breve{\mathbb{M}}_j(\mathbf{t})\big)^{\frac{1}{2}}. \tag{7.48}$$

Therefore, what remains is to estimate the integral in (7.48), which can be done by elementary Gaussian integral step by step. More specifically, using (7.44) and (7.45) and the change of variable $t_j/t_{j-1} - 1 \to t_j$ in case of $t_{j-1} \in \mathbb{I}^c$ and $t_j - t_{j-1} \to t_j$ in case of $t_{j-1} \in \mathbb{I}$, it is elementary to see that for any $\ell = O(W)$,

$$\int_{\mathbb{R}_+} dt_j \, t_j^\ell \big(\breve{\mathbb{M}}_j(\mathbf{t})\big)^{\frac{1}{2}} \le \ell!!\Big(1 + c\mathfrak{L}^{-\frac{1}{2}}\Big)^{O(\ell)} \big(t_{j-1}^{\ell+1} + 1\big) \le e^{O(W\log N)}\Big(1 + \mathfrak{L}^{-\frac{1}{2}}\Big)^{O(\ell)} \big(t_{j-1}^{\ell+1} + 1\big). \tag{7.49}$$

Starting from $j = W$, using (7.49) to integrate (7.48) successively, the exponent of t_j increases linearly ($n_j = O(1)$), thus we can get

$$\int_{\prod_{j=2}^W \mathbb{J}_j} \prod_{j=2}^W \mathrm{d}t_j \, \mathbb{M}(\mathbf{t})\mathfrak{q}(\mathbf{t}) \le e^{-\frac{\mathfrak{L}}{600W^2}} \cdot e^{O(W^2\log N)} \cdot \Big(1 + \mathfrak{L}^{-\frac{1}{2}}\Big)^{O(W^2)}.$$

Then (7.4) follows from the definition of $\mathfrak{L}$ in (7.3) and (5.38). Hence, we completed the proof for (7.4) when the spanning tree is given by (7.40).

Now, we consider more general spanning tree $\mathcal{G}_0$ and regard 1 as its root. We start from the generalization of (7.41), namely,

$$\mathbb{M}(\mathbf{t}) \leq \prod_{\{i,j\}\in\mathcal{E}_0} \exp\Big\{ - \mathfrak{L}\frac{(t_j - t_i)^2}{1+2t_i^2}\Big\} := \prod_{\{i,j\}\in\mathcal{E}_0} \breve{\mathbb{M}}_{i,j}(\mathbf{t}). \tag{7.50}$$

Here we make the convention that $\mathrm{dist}(1,i) = \mathrm{dist}(1,j) - 1$ for all $\{i,j\} \in \mathcal{E}_0$, where $\mathrm{dist}(a,b)$ represents the distance between a and b. Now, if there is k' such that $\mathbb{J}_{k'} \in \mathbb{I}^c$, we can prove the following analogue of (7.43), namely,

$$\sum_{\{i,j\}\in\mathcal{E}_0} \frac{(t_j - t_i)^2}{1+2t_i^2} \geq \frac{1}{300k^2}$$

by performing the argument in (7.44)-(7.47) on the path connecting k' and the root 1. Consequently, we can get the analogue of (7.48) via replacing $\breve{\mathbb{M}}_j(\mathbf{t})$'s by $\breve{\mathbb{M}}_{i,j}(\mathbf{t})$'s. Finally, integrating t_j's out successively, from the leaves to the root 1, yields the same conclusion, i.e. (7.4), for general $\mathcal{G}_0$. Therefore, we completed the proof of Lemma 7.2.

8. Gaussian measure in the vicinities

From now on, we can restrict ourselves to the Type I, II and III vicinities. As a preparation of the proofs of Lemmas 5.8 and 5.9, we will show in this section that the exponential function

$$\exp\Big\{-M\big(\mathring{K}(\hat{X},V) + \mathring{L}(\hat{B},T)\big)\Big\} \tag{8.1}$$

is approximately a Gaussian measure (unnormalized).

8.1. **Parametrization and initial approximation in the vicinities.** We change the $\mathbf{x}$, $\mathbf{b}$, $\mathbf{t}$, $\mathbf{v}$-variables to a new set of variables, namely, $\mathring{\mathbf{x}}$, $\mathring{\mathbf{b}}$, $\mathring{\mathbf{t}}$ and $\mathring{\mathbf{v}}$. The precise definition of $\mathring{\hat{x}}$ differs in the different vicinities. To distinguish the parameterization, we set $\varkappa = \pm$, $+$, or $-$, corresponding to Type I, II or III vicinity, respectively. Recalling $D_\varkappa$ from (1.24). For each j and each $\varkappa$, we then set

$$\begin{aligned} &\hat{X}_j = D_\varkappa \mathrm{diag}\Big(\exp\big\{\mathbf{i}\mathring{x}_{j,1}/\sqrt{M}\big\}, \exp\big\{\mathbf{i}\mathring{x}_{j,2}/\sqrt{M}\big\}\Big), \quad \mathring{x}_{j,a}/\sqrt{M} \in [-\pi,\pi], \\ &\hat{B}_j = D_\pm + D_\pm \mathrm{diag}\Big(\mathring{b}_{j,1}/\sqrt{M}, \mathring{b}_{j,2}/\sqrt{M}\Big), \qquad t_j = \mathring{t}_j/\sqrt{M}. \end{aligned} \tag{8.2}$$

If $\varkappa = \pm$, we also need to parameterize v_j by

$$v_j = \mathring{v}_j/\sqrt{M}. \tag{8.3}$$

We set the vectors

$$\begin{aligned} &\mathring{\mathbf{b}}_a := (\mathring{b}_{1,a}, \ldots, \mathring{b}_{W,a}), \quad \mathring{\mathbf{x}}_a := (\mathring{x}_{1,a}, \ldots, \mathring{x}_{W,a}), \qquad a = 1,2, \\ &\mathring{\mathbf{t}} := (\mathring{t}_2, \ldots, \mathring{t}_W), \quad \mathring{\mathbf{v}} := (\mathring{v}_2, \ldots, \mathring{v}_W). \end{aligned}$$

Accordingly, recalling the quantity Θ from (5.33), we introduce the domains

$$\mathring{\Upsilon} \equiv \mathring{\Upsilon}(N,\varepsilon_0) := \{\mathbf{a} \in \mathbb{R}^W : ||\mathbf{a}||_2^2 \leq \Theta\}, \quad \mathring{\Upsilon}_S \equiv \mathring{\Upsilon}_S(N,\varepsilon_0) := \{\mathbf{a} \in \mathbb{R}_+^{W-1} : -\mathbf{a}' S^{(1)} \mathbf{a} \leq \Theta\}.$$

We remind here, as mentioned above, in the sequel, the small constant ε_0 in $\mathring{\Upsilon}$ and $\mathring{\Upsilon}_S$ may be different from line to line, subject to (5.37). Now, by the definition of the Type I', II and III vicinities in Definition 5.5 and the parametrization in (8.2) and (8.3), we can redefine the vicinities as follows.

Definition 8.1. *We can redefine three types of vicinities as follows.*

- *Type I' vicinity :* $\big(\mathring{\mathbf{b}}_1, \mathring{\mathbf{b}}_2, \mathring{\mathbf{x}}_1, \mathring{\mathbf{x}}_2, \mathring{\mathbf{t}}, \mathring{\mathbf{v}}\big) \in \mathring{\Upsilon} \times \mathring{\Upsilon} \times \mathring{\Upsilon} \times \mathring{\Upsilon} \times \mathring{\Upsilon}_S \times \mathring{\Upsilon}_S$, *with* $\varkappa = \pm$.
- *Type II vicinity :* $\big(\mathring{\mathbf{b}}_1, \mathring{\mathbf{b}}_2, \mathring{\mathbf{x}}_1, \mathring{\mathbf{x}}_2, \mathring{\mathbf{t}}, \mathbf{v}\big) \in \mathring{\Upsilon} \times \mathring{\Upsilon} \times \mathring{\Upsilon} \times \mathring{\Upsilon} \times \mathring{\Upsilon}_S \times \mathbb{I}^{W-1}$, *with* $\varkappa = +$.
- *Type III vicinity :* $\big(\mathring{\mathbf{b}}_1, \mathring{\mathbf{b}}_2, \mathring{\mathbf{x}}_1, \mathring{\mathbf{x}}_2, \mathring{\mathbf{t}}, \mathbf{v}\big) \in \mathring{\Upsilon} \times \mathring{\Upsilon} \times \mathring{\Upsilon} \times \mathring{\Upsilon} \times \mathring{\Upsilon}_S \times \mathbb{I}^{W-1}$, *with* $\varkappa = -$.

We recall from (7.6) the fact

$$\mathring{\mathbf{t}} \in \mathring{\Upsilon}_S \Longrightarrow ||\mathring{\mathbf{t}}||_\infty = O(\Theta). \tag{8.4}$$

Now, we use the representation (5.2). Then, for the Type I vicinity, we change $\mathbf{x}, \mathbf{b}, \mathbf{t}, \mathbf{v}$-variables to $\mathring{\mathbf{x}}, \mathring{\mathbf{b}}, \mathring{\mathbf{t}}, \mathring{\mathbf{v}}$-variables according to (8.2) with $\varkappa = \pm$, thus

$$\begin{aligned}
2^W \mathcal{I}\big(\Upsilon_+^b, \Upsilon_-^b, \Upsilon_+^x, \Upsilon_-^x, \Upsilon_S, \Upsilon_S\big) = \frac{M^2}{(n!)^2 4^W \pi^{2W+4}} \int_{\mathbb{L}^{2W-2}} \prod_{j=2}^W \frac{\mathrm{d}\theta_j}{2\pi} \prod_{j=2}^W \frac{\mathrm{d}\sigma_j}{2\pi} \int_{\mathring{\Upsilon}} \prod_{j=1}^W \mathrm{d}\mathring{b}_{j,1} \int_{\mathring{\Upsilon}} \prod_{j=1}^W \mathrm{d}\mathring{b}_{j,2} \\
\times \int_{\mathring{\Upsilon}} \prod_{j=1}^W \mathrm{d}\mathring{x}_{j,1} \int_{\mathring{\Upsilon}} \prod_{j=1}^W \mathrm{d}\mathring{x}_{j,2} \int_{\mathring{\Upsilon}_S} \prod_{j=2}^W 2\mathring{t}_j \mathrm{d}\mathring{t}_j \int_{\mathring{\Upsilon}_S} \prod_{j=2}^W 2\mathring{v}_j \mathrm{d}\mathring{v}_j \prod_{j=1}^W \exp\Big\{\mathbf{i}\frac{\mathring{x}_{j,1} + \mathring{x}_{j,2}}{\sqrt{M}}\Big\} \\
\times \exp\Big\{ - M\big(\mathring{K}(\hat{X}, V) + \mathring{L}(\hat{B}, T)\big)\Big\} \cdot \prod_{j=1}^W (x_{j,1} - x_{j,2})^2 (b_{j,1} + b_{j,2})^2 \cdot \mathsf{A}(\hat{X}, \hat{B}, V, T).
\end{aligned} \tag{8.5}$$

For the Type II or III vicinities, i.e. $\varkappa = +$ or $-$, we change $\mathbf{x}, \mathbf{b}, \mathbf{t}$-variables to $\mathring{\mathbf{x}}, \mathring{\mathbf{b}}, \mathring{\mathbf{t}}$-variables. Consequently, we have

$$\begin{aligned}
\mathcal{I}\big(\Upsilon_+^b, \Upsilon_-^b, \Upsilon_\varkappa^x, \Upsilon_\varkappa^x, \Upsilon_S, \mathbb{I}^{W-1}\big) = \frac{(-a_\varkappa^2)^W}{(n!)^2} \cdot \frac{M^{W+1}}{8^W \pi^{2W+4}} \cdot \int_{\mathbb{L}^{2W-2}} \prod_{j=2}^W \frac{\mathrm{d}\theta_j}{2\pi} \prod_{j=2}^W \frac{\mathrm{d}\sigma_j}{2\pi} \int_{\mathring{\Upsilon}} \prod_{j=1}^W \mathrm{d}\mathring{b}_{j,1} \int_{\mathring{\Upsilon}} \prod_{j=1}^W \mathrm{d}\mathring{b}_{j,2} \\
\times \int_{\mathring{\Upsilon}} \prod_{j=1}^W \mathrm{d}\mathring{x}_{j,1} \int_{\mathring{\Upsilon}} \prod_{j=1}^W \mathrm{d}\mathring{x}_{j,2} \int_{\mathring{\Upsilon}_S} \prod_{j=2}^W 2\mathring{t}_j \mathrm{d}\mathring{t}_j \int_{\mathbb{I}^{W-1}} \prod_{j=2}^W 2v_j \mathrm{d}v_j \prod_{j=1}^W \exp\Big\{\mathbf{i}\frac{\mathring{x}_{j,1} + \mathring{x}_{j,2}}{\sqrt{M}}\Big\} \\
\times \exp\Big\{ - M\big(\mathring{K}(\hat{X}, V) + \mathring{L}(\hat{B}, T)\big)\Big\} \cdot \prod_{j=1}^W (x_{j,1} - x_{j,2})^2 (b_{j,1} + b_{j,2})^2 \cdot \mathsf{A}(\hat{X}, \hat{B}, V, T).
\end{aligned} \tag{8.6}$$

We will also need the following facts

$$\prod_{j=1}^W |(x_{j,1} - x_{j,2})^2 (b_{j,1} + b_{j,2})^2| = e^{O(W)}, \qquad |\mathsf{A}(\hat{X}, \hat{B}, V, T)| \le e^{O(WN^{\varepsilon_2})} \tag{8.7}$$

if

$$\mathbf{x}_1, \mathbf{x}_2 \in \widehat{\Sigma}^W, \quad b_{j,1} = a_+ + o(1), \quad b_{j,2} = -a_- + o(1), \quad t_j = o(1), \quad \forall\, j = 1, \dots, N,$$

which always hold in these types of vicinities. The first estimate in (8.7) is trivial, and the second follows from Lemma 6.1.

Now, we approximate (8.1) in the vicinities. For any $\vartheta \in \mathbb{L}$, we introduce the matrices

$$\mathcal{E}_+(\vartheta) := \begin{pmatrix} 0 & e^{\mathbf{i}\vartheta} \\ e^{-\mathbf{i}\vartheta} & 0 \end{pmatrix}, \qquad \mathcal{E}_-(\vartheta) := \begin{pmatrix} 0 & e^{\mathbf{i}\vartheta} \\ -e^{-\mathbf{i}\vartheta} & 0 \end{pmatrix}.$$

Then, with the parameterization above, expanding $\hat{X}_j$ in (3.22) and T_j in (3.24) up to the second order, we can write

$$\begin{aligned}
\hat{X}_j &= D_\varkappa + \frac{\mathbf{i}}{\sqrt{M}} D_\varkappa \mathrm{diag}(\mathring{x}_{j,1}, \mathring{x}_{j,2}) + \frac{1}{M} R_j^x, \quad \varkappa = \pm, +, -, \\
T_j &= I + \frac{\mathring{t}_j}{\sqrt{M}} \mathcal{E}_+(\sigma_j) + \frac{1}{M} R_j^t.
\end{aligned} \tag{8.8}$$

For $\varkappa = \pm$, we also expand V_j in (3.24) up to the second order, namely,

$$V_j = I + \frac{\mathring{v}_j}{\sqrt{M}} \mathcal{E}_-(\theta_j) + \frac{1}{M} R_j^v. \tag{8.9}$$

We just take (8.8) and (8.9) as the definition of R^x_j, R^t_j and R^v_j. Note that R^x_j is actually $\varkappa$-dependent. However, this dependence is irrelevant for our analysis thus is suppressed from the notation. It is elementary that

$$||R^x_j||_{\max} = O(\mathring{x}^2_{j,1} + \mathring{x}^2_{j,2}), \quad ||R^t_j||_{\max} = O(\mathring{t}^2_j), \quad ||R^v_j||_{\max} = O(\mathring{v}^2_j). \tag{8.10}$$

Here $||\cdot||_{\max}$ represents the max-norm of a matrix.

Recall the facts (5.10) and (5.19)

$$\begin{aligned} M\mathring{L}(\hat{B},T) &= M\Big(\mathring{\ell}_{++}(\mathbf{b}_1) + \mathring{\ell}_{--}(\mathbf{b}_2)\Big) + M\ell_S(\hat{B},T), \\ M\mathring{K}(\hat{X},V) &= M\Big(-\mathring{\ell}_{++}(\mathbf{x}_1) - \mathring{\ell}_{+-}(\mathbf{x}_2)\Big) + M\ell_S(\hat{X},V). \end{aligned} \tag{8.11}$$

In light of (5.17)-(5.19), we can also represent $M\mathring{K}(\hat{X},V)$ in the following two alternative ways

$$M\mathring{K}(\hat{X},V) = M\Big(-\mathring{\ell}_{++}(\mathbf{x}_1) - \mathring{\ell}_{++}(\mathbf{x}_2)\Big) + M\ell_S(\hat{X},V) + M\big(K(D_+,I) - K(D_\pm,I)\big), \tag{8.12}$$

$$M\mathring{K}(\hat{X},V) = M\Big(-\mathring{\ell}_{+-}(\mathbf{x}_1) - \mathring{\ell}_{+-}(\mathbf{x}_2)\Big) + M\ell_S(\hat{X},V) + M\big(K(D_-,I) - K(D_\pm,I)\big). \tag{8.13}$$

We will use three representations of $M\mathring{K}(\hat{X},V)$ in (8.11), (8.12) and (8.13) for Type I', II and III vicinities respectively. In addition, we introduce the matrices

$$\mathbb{A}_+ := (1+a_+^2)I + a_+^2 S, \qquad \mathbb{A}_- := (1+a_-^2)I + a_-^2 S. \tag{8.14}$$

Then, we have the following lemma.

Lemma 8.2. *With the parametrization in (8.8), we have the following approximations.*

- *Let $\mathring{\mathbf{b}}_1, \mathring{\mathbf{b}}_2 \in \mathbb{C}^W$ and $||\mathring{\mathbf{b}}_1||_\infty, ||\mathring{\mathbf{b}}_2||_\infty = o(\sqrt{M})$, we have*

$$M\Big(\mathring{\ell}_{++}(\mathbf{b}_1) + \mathring{\ell}_{--}(\mathbf{b}_2)\Big) = \frac{1}{2}\mathring{\mathbf{b}}_1'\mathbb{A}_+\mathring{\mathbf{b}}_1 + \frac{1}{2}\mathring{\mathbf{b}}_2'\mathbb{A}_-\mathring{\mathbf{b}}_2 + R^b, \quad R^b = O\Big(\frac{\sum_{a=1,2}||\mathring{\mathbf{b}}_a||_3^3}{\sqrt{M}}\Big). \tag{8.15}$$

- *Let $\varkappa = \pm$ and $\mathring{\mathbf{x}}_1, \mathring{\mathbf{x}}_2 \in \mathbb{C}^W$ and $||\mathring{\mathbf{x}}_1||_\infty, ||\mathring{\mathbf{x}}_2||_\infty = o(\sqrt{M})$, we have*

$$M\Big(-\mathring{\ell}_{++}(\mathbf{x}_1) - \mathring{\ell}_{+-}(\mathbf{x}_2)\Big) = \frac{1}{2}\mathring{\mathbf{x}}_1'\mathbb{A}_+\mathring{\mathbf{x}}_1 + \frac{1}{2}\mathring{\mathbf{x}}_2'\mathbb{A}_-\mathring{\mathbf{x}}_2 + R^x_\pm, \quad R^x_\pm = O\Big(\frac{\sum_{a=1,2}||\mathring{\mathbf{x}}_a||_3^3}{\sqrt{M}}\Big). \tag{8.16}$$

- *In the Type II vicinity, we have*

$$M\Big(-\mathring{\ell}_{++}(\mathbf{x}_1) - \mathring{\ell}_{++}(\mathbf{x}_2)\Big) = \frac{1}{2}\mathring{\mathbf{x}}_1'\mathbb{A}_+\mathring{\mathbf{x}}_1 + \frac{1}{2}\mathring{\mathbf{x}}_2'\mathbb{A}_+\mathring{\mathbf{x}}_2 + R^x_+, \quad R^x_+ = O\Big(\frac{\Theta^{\frac{3}{2}}}{\sqrt{M}}\Big). \tag{8.17}$$

- *In the Type III vicinity, we have*

$$M\Big(-\mathring{\ell}_{+-}(\mathbf{x}_1) - \mathring{\ell}_{+-}(\mathbf{x}_2)\Big) = \frac{1}{2}\mathring{\mathbf{x}}_1'\mathbb{A}_-\mathring{\mathbf{x}}_1 + \frac{1}{2}\mathring{\mathbf{x}}_2'\mathbb{A}_-\mathring{\mathbf{x}}_2 + R^x_-, \quad R^x_- = O\Big(\frac{\Theta^{\frac{3}{2}}}{\sqrt{M}}\Big). \tag{8.18}$$

Here R^b $R^x_\pm$, R^x_+ and R^x_- are remainder terms of the Taylor expansion of the function $\ell(\mathbf{a})$ defined in (5.6).

Remark 8.3. *Here we stated (8.15) and (8.16) in the domains much larger than the Type I' vicinity for further discussion. In addition, the restriction $||\mathring{\mathbf{b}}_a||_\infty$ and $||\mathring{\mathbf{x}}_a||_\infty$ for $a = 1, 2$ is imposed to avoid the ambiguity of the definition of the logarithmic term in the function $\ell(\mathbf{a})$.*

Proof. It follows from the Taylor expansion of the function $\ell(\mathbf{a})$ easily. □

Then, according to (8.11)-(8.13), what remains is to approximate $M\ell_S(\hat{B},T)$ and $M\ell_S(\hat{X},V)$ in the vicinities. Recalling the definition in (5.6) and the parameterization in (8.2), we can rewrite

$$\begin{aligned} M\ell_S(\hat{B},T) &= \frac{1}{2}\sum_{j,k}\mathfrak{s}_{jk}|s_j\mathring{t}_k e^{\mathbf{i}\sigma_k} - s_k\mathring{t}_j e^{\mathbf{i}\sigma_j}|^2 \\ &\quad\times\Big(a_+ - a_- + \frac{a_+\mathring{b}_{j,1} - a_-\mathring{b}_{j,2}}{\sqrt{M}}\Big)\Big(a_+ - a_- + \frac{a_+\mathring{b}_{k,1} - a_-\mathring{b}_{k,2}}{\sqrt{M}}\Big) \\ &=: \frac{(a_+ - a_-)^2}{2}\sum_{j,k}\mathfrak{s}_{jk}|\mathring{t}_k e^{\mathbf{i}\sigma_k} - \mathring{t}_j e^{\mathbf{i}\sigma_j}|^2 + R^{t,b}. \end{aligned} \tag{8.19}$$

We take the above equation as the definition of $R^{t,b}$. Now, we set

$$\tau_{j,1} := \mathring{t}_j\cos\sigma_j, \qquad \tau_{j,2} := \mathring{t}_j\sin\sigma_j, \quad \forall\, j = 2,\dots,W$$

and change the variables and the measure as

$$(\mathring{t}_j,\sigma_j)\to(\tau_{j,1},\tau_{j,2}), \qquad 2\mathring{t}_j\mathrm{d}\mathring{t}_j\frac{\mathrm{d}\sigma_j}{2\pi}\to\frac{1}{\pi}\mathrm{d}\tau_{j,1}\mathrm{d}\tau_{j,2}. \tag{8.20}$$

In the Type I' vicinity, we can do the same thing for $M\ell_S(\hat{X},V)$, namely,

$$M\ell_S(\hat{X},V) =: \frac{(a_+ - a_-)^2}{2}\sum_{j,k}\mathfrak{s}_{jk}|\mathring{v}_k e^{\mathbf{i}\theta_k} - \mathring{v}_j e^{\mathbf{i}\theta_j}|^2 + R_{\pm}^{v,x}, \tag{8.21}$$

where $R_{\pm}^{v,x}$ is the remainder term. Then we set

$$\upsilon_{j,1} := \mathring{v}_j\cos\theta_j, \qquad \upsilon_{j,2} := \mathring{v}_j\sin\theta_j, \quad \forall\, j = 2,\dots,W$$

and change the variables and measure as

$$(\mathring{v}_j,\theta_j)\to(\upsilon_{j,1},\upsilon_{j,2}), \qquad 2\mathring{v}_j\mathrm{d}\mathring{v}_j\frac{\mathrm{d}\theta_j}{2\pi}\to\frac{1}{\pi}\mathrm{d}\upsilon_{j,1}\mathrm{d}\upsilon_{j,2}. \tag{8.22}$$

Now, we introduce the vectors

$$\boldsymbol{\tau}_a = (\tau_{2,a},\dots,\tau_{W,a}), \qquad \boldsymbol{\upsilon}_a = (\upsilon_{2,a},\dots,\upsilon_{W,a}), \qquad a = 1,2.$$

With this notation, we can rewrite (8.19) and (8.21) as

$$\begin{aligned} M\ell_S(\hat{B},T) &= -(a_+ - a_-)^2\sum_{a=1,2}\boldsymbol{\tau}_a' S^{(1)}\boldsymbol{\tau}_a + R^{t,b}, \\ M\ell_S(\hat{X},V) &= -(a_+ - a_-)^2\sum_{a=1,2}\boldsymbol{\upsilon}_a' S^{(1)}\boldsymbol{\upsilon}_a + R_{\pm}^{v,x}. \end{aligned} \tag{8.23}$$

According to (8.20) and (8.22), we can express (8.5) as an integral over $\mathring{\mathbf{b}}$, $\mathring{\mathbf{x}}$, $\mathring{\boldsymbol{\tau}}$ and $\mathring{\boldsymbol{\upsilon}}$-variables. However, we need to specify the domains of $\mathring{\boldsymbol{\tau}}$ and $\mathring{\boldsymbol{\upsilon}}$-variables in advance. Our aim is to restrict the integral in the domains

$$\boldsymbol{\tau}_a\in\mathring{\Upsilon}_S, \quad \boldsymbol{\upsilon}_a\in\mathring{\Upsilon}_S, \qquad a = 1,2. \tag{8.24}$$

Taking $\mathring{\mathbf{t}}$ for instance, we see that

$$(t_j - t_k)^2 \le |t_j e^{\mathbf{i}\sigma_j} - t_k e^{\mathbf{i}\sigma_k}|^2 = (\tau_{j,1} - \tau_{k,1})^2 + (\tau_{j,2} - \tau_{k,2})^2,$$

which actually implies

$$\boldsymbol{\tau}_a\in\mathring{\Upsilon}_S \quad \text{for} \quad a = 1,2 \Longrightarrow \mathring{\mathbf{t}}\in\mathring{\Upsilon}_S. \tag{8.25}$$

However the reverse of (8.25) may not be true. That means, (8.24) is stronger than $(\mathring{\mathbf{t}},\mathring{\mathbf{v}},\boldsymbol{\sigma},\boldsymbol{\theta})\in\mathring{\Upsilon}_S\times\mathring{\Upsilon}_S\times\mathbb{L}^{W-1}\times\mathbb{L}^{W-1}$. To show the truncation to (8.24) from $(\mathring{\mathbf{t}},\mathring{\mathbf{v}},\boldsymbol{\sigma},\boldsymbol{\theta})\in\mathring{\Upsilon}_S\times\mathring{\Upsilon}_S\times\mathbb{L}^{W-1}\times\mathbb{L}^{W-1}$

is harmless in the integral (8.5), we need to bound $R^{t,b}$ and $R^{v,x}_{\pm}$ in terms of $\boldsymbol{\tau}_a' S^{(1)} \boldsymbol{\tau}_a$ and $\boldsymbol{\upsilon}_a' S^{(1)} \boldsymbol{\upsilon}_a$, respectively. More specifically, we need the following lemma.

Lemma 8.4. *In the Type I' vicinity, we have*

$$|R^{t,b}| \le O\Big(\frac{\Theta^{\frac{1}{2}}}{\sqrt{M}}\Big) \sum_{a=1,2} \big(- \boldsymbol{\tau}_a' S^{(1)} \boldsymbol{\tau}_a\big) + O\Big(\frac{\Theta^{\frac{7}{2}}}{M}\Big) \sum_{a=1,2} \big(- \boldsymbol{\tau}_a' S^{(1)} \boldsymbol{\tau}_a\big)^{1/2} + O\Big(\frac{\Theta^7}{M^2}\Big),$$
$$|R^{v,x}_{\pm}| \le O\Big(\frac{\Theta^{\frac{1}{2}}}{\sqrt{M}}\Big) \sum_{a=1,2} \big(- \boldsymbol{\upsilon}_a' S^{(1)} \boldsymbol{\upsilon}_a\big) + O\Big(\frac{\Theta^{\frac{7}{2}}}{M}\Big) \sum_{a=1,2} \big(- \boldsymbol{\upsilon}_a' S^{(1)} \boldsymbol{\upsilon}_a\big)^{1/2} + O\Big(\frac{\Theta^7}{M^2}\Big). \tag{8.26}$$

Proof. Since the proofs of these two bounds are nearly the same, we only state the details for the first one. By (8.4), we have

$$s_j = 1 + O\Big(\frac{\Theta^2}{M}\Big), \quad \forall\, j = 2, \dots, W.$$

Then it is not difficult to see that

$$\begin{aligned} |s_j \mathring{t}_k e^{\mathbf{i}\sigma_k} - s_k \mathring{t}_j e^{\mathbf{i}\sigma_j}|^2 &= \Big|\mathring{t}_k e^{\mathbf{i}\sigma_k} - \mathring{t}_j e^{\mathbf{i}\sigma_j} + O\Big(\frac{\Theta^3}{M}\Big)\Big|^2 \\ &= \sum_{a=1,2} (\tau_{j,a} - \tau_{k,a})^2 + O\Big(\frac{\Theta^3}{M}\Big) \sum_{a=1,2} |\tau_{j,a} - \tau_{k,a}| + O\Big(\frac{\Theta^6}{M^2}\Big). \end{aligned} \tag{8.27}$$

Now, by the fact from Definition 8.1

$$||\mathring{\mathbf{b}}_a||_\infty \le ||\mathring{\mathbf{b}}_a||_2 = O(\Theta^{\frac{1}{2}}) \quad \text{for} \quad a = 1, 2, \tag{8.28}$$

we have

$$\Big(a_+ - a_- + \frac{a_+ \mathring{b}_{j,1} - a_- \mathring{b}_{j,2}}{\sqrt{M}}\Big)\Big(a_+ - a_- + \frac{a_+ \mathring{b}_{k,1} - a_- \mathring{b}_{k,2}}{\sqrt{M}}\Big) = (a_+ - a_-)^2 + O\Big(\frac{\Theta^{\frac{1}{2}}}{\sqrt{M}}\Big). \tag{8.29}$$

Substituting (8.27) and (8.29) into (8.19) yields

$$\begin{aligned} |R^{t,b}| &\le O\Big(\frac{\Theta^{\frac{1}{2}}}{\sqrt{M}}\Big) \sum_{a=1,2} \sum_{j,k} \mathfrak{s}_{jk} (\tau_{j,a} - \tau_{k,a})^2 + O\Big(\frac{\Theta^3}{M}\Big) \sum_{a=1,2} \sum_{j,k} \mathfrak{s}_{jk} |\tau_{j,a} - \tau_{k,a}| + O\Big(\frac{\Theta^7}{M^2}\Big) \\ &\le O\Big(\frac{\Theta^{\frac{1}{2}}}{\sqrt{M}}\Big) \sum_{a=1,2} \big(- \boldsymbol{\tau}_a' S^{(1)} \boldsymbol{\tau}_a\big) + O\Big(\frac{\Theta^{\frac{7}{2}}}{M}\Big) \sum_{a=1,2} \big(- \boldsymbol{\tau}_a' S^{(1)} \boldsymbol{\tau}_a\big)^{1/2} + O\Big(\frac{\Theta^7}{M^2}\Big). \end{aligned}$$

Here we used Cauchy Schwarz inequality and $\sum_{j,k} \mathfrak{s}_{jk} = O(W) = O(\Theta)$ in the second step. The bound for $|R^{v,x}_{\pm}|$ can be proved analogously. Hence, we completed the proof of Lemma 8.4. □

Roughly speaking, by (8.23) and Lemma 8.4 we have

$$\begin{aligned} M\ell_S(\hat{B}, T) &= -\big(a_+ - a_- + o(1)\big)^2 \sum_{a=1,2} \boldsymbol{\tau}_a' S^{(1)} \boldsymbol{\tau}_a + o\Big(\sum_{a=1,2} \big(- \boldsymbol{\tau}_a' S^{(1)} \boldsymbol{\tau}_a\big)^{1/2}\Big) + o(1), \\ M\ell_S(\hat{X}, V) &= -\big(a_+ - a_- + o(1)\big)^2 \sum_{a=1,2} \boldsymbol{\upsilon}_a' S^{(1)} \boldsymbol{\upsilon}_a + o\Big(\sum_{a=1,2} \big(- \boldsymbol{\upsilon}_a' S^{(1)} \boldsymbol{\upsilon}_a\big)^{1/2}\Big) + o(1). \end{aligned}$$

Then it is obvious that if one of $\boldsymbol{\tau}_1$, $\boldsymbol{\tau}_2$, $\boldsymbol{\upsilon}_1$ and $\boldsymbol{\upsilon}_2$ is not in $\mathring{\Upsilon}_S$, we will get (7.26). Hence, using (8.7), we can discard the integral outside the vicinity, analogously to the proof of Lemma 7.7. More specifically, in the sequel, we can and do assume

$$\boldsymbol{\tau}_1, \boldsymbol{\tau}_2, \boldsymbol{\upsilon}_1, \boldsymbol{\upsilon}_2 \in \mathring{\Upsilon}_S. \tag{8.30}$$

Now, plugging (8.30) into (8.26) in turn yields the bound

$$|R^{t,b}| = O\Big(\frac{\Theta^{\frac{3}{2}}}{\sqrt{M}} \vee \frac{\Theta^4}{M}\Big), \quad |R^{v,x}_{\pm}| = O\Big(\frac{\Theta^{\frac{3}{2}}}{\sqrt{M}} \vee \frac{\Theta^4}{M}\Big). \tag{8.31}$$

By the discussion above, for the Type I vicinity, we can write (8.5) as

$$
\begin{aligned}
2^W\mathcal{I}(\Upsilon_+^b,\Upsilon_-^b,\Upsilon_+^x,\Upsilon_-^x,\Upsilon_S,\Upsilon_S) &= \frac{M^2}{(n!)^2 4^W\pi^{4W+2}}\cdot\int_{\mathring{\Upsilon}}\prod_{j=1}^W \mathrm{d}\mathring{b}_{j,1}\int_{\mathring{\Upsilon}}\prod_{j=1}^W \mathrm{d}\mathring{b}_{j,2}\\
&\times\int_{\mathring{\Upsilon}}\prod_{j=1}^W \mathrm{d}\mathring{x}_{j,1}\int_{\mathring{\Upsilon}}\prod_{j=1}^W \mathrm{d}\mathring{x}_{j,2}\int_{\mathring{\Upsilon}_S}\prod_{j=2}^W \mathrm{d}\tau_{j,1}\int_{\mathring{\Upsilon}_S}\prod_{j=2}^W \mathrm{d}\tau_{j,2}\int_{\mathring{\Upsilon}_S}\prod_{j=2}^W \mathrm{d}\upsilon_{j,1}\int_{\mathring{\Upsilon}_S}\prod_{j=2}^W \mathrm{d}\upsilon_{j,2}\\
&\times\exp\Big\{-\frac{1}{2}\mathring{\mathbf{b}}_1'\mathbb{A}_+\mathring{\mathbf{b}}_1-\frac{1}{2}\mathring{\mathbf{b}}_2'\mathbb{A}_-\mathring{\mathbf{b}}_2-R^b\Big\}\cdot\exp\Big\{-\frac{1}{2}\mathring{\mathbf{x}}_1'\mathbb{A}_+\mathring{\mathbf{x}}_1-\frac{1}{2}\mathring{\mathbf{x}}_2'\mathbb{A}_-\mathring{\mathbf{x}}_2-R_\pm^x\Big\}\\
&\times\exp\Big\{(a_+-a_-)^2\sum_{a=1,2}\boldsymbol{\tau}_a'S^{(1)}\boldsymbol{\tau}_a-R^{t,b}\Big\}\cdot\exp\Big\{(a_+-a_-)^2\sum_{a=1,2}\boldsymbol{\upsilon}_a'S^{(1)}\boldsymbol{\upsilon}_a-R_\pm^{v,x}\Big\}\\
&\times\prod_{j=1}^W\exp\Big\{\mathbf{i}\frac{\mathring{x}_{j,1}+\mathring{x}_{j,2}}{\sqrt{M}}\Big\}\cdot\prod_{j=1}^W(x_{j,1}-x_{j,2})^2(b_{j,1}+b_{j,2})^2\cdot\mathsf{A}(\hat{X},\hat{B},V,T)+O(e^{-\Theta}),
\end{aligned}
\tag{8.32}
$$

where the error term stems from the truncation of the vicinity $(\mathring{\mathbf{t}},\mathring{\mathbf{v}},\boldsymbol{\sigma},\boldsymbol{\theta})\in\mathring{\Upsilon}_S\times\mathring{\Upsilon}_S\times\mathbb{L}^{W-1}\times\mathbb{L}^{W-1}$ to $(\boldsymbol{\tau}_1,\boldsymbol{\tau}_2,\boldsymbol{\upsilon}_1,\boldsymbol{\upsilon}_2)\in\mathring{\Upsilon}_S\times\mathring{\Upsilon}_S\times\mathring{\Upsilon}_S\times\mathring{\Upsilon}_S$.

Now, for the Type II and III vicinities, the discussion on $\ell_S(\hat{B},T)$ is of course the same. For $\ell_S(\hat{X},V)$, we make the following approximation. For the Type II vicinity, using the notation in (5.21),we can write

$$
\begin{aligned}
M\ell_S(\hat{X},V) &= \frac{-Ma_+^2}{2}\sum_{j,k}\mathfrak{s}_{jk}^v\Big(\frac{\mathring{x}_{j,1}}{\sqrt{M}}-\frac{\mathring{x}_{j,2}}{\sqrt{M}}+O\Big(\frac{\mathring{x}_{j,1}^2+\mathring{x}_{j,2}^2}{M}\Big)\Big)\Big(\frac{\mathring{x}_{k,1}}{\sqrt{M}}-\frac{\mathring{x}_{k,2}}{\sqrt{M}}+O\Big(\frac{\mathring{x}_{k,1}^2+\mathring{x}_{k,2}^2}{M}\Big)\Big)\\
&=: -\frac{a_+^2}{2}\sum_{j,k}\mathfrak{s}_{jk}^v(\mathring{x}_{j,1}-\mathring{x}_{j,2})(\mathring{x}_{k,1}-\mathring{x}_{k,2})+R_+^{v,x}.
\end{aligned}
\tag{8.33}
$$

It is easy to see that

$$
R_+^{v,x}=O\Big(\frac{||\mathring{\mathbf{x}}_1||_3^3+||\mathring{\mathbf{x}}_2||_3^3}{\sqrt{M}}\Big)=O\Big(\frac{\Theta^{\frac{3}{2}}}{\sqrt{M}}\Big),\quad\text{for}\quad\mathring{\mathbf{x}}_1,\mathring{\mathbf{x}}_2\in\mathring{\Upsilon}.
\tag{8.34}
$$

Combining (5.15), (5.18), (8.17) and (8.33) we obtain

$$
\begin{aligned}
M(K(\hat{X},V)-K(D_+,I)) &= \frac{1}{2}\mathring{\mathbf{x}}_1'\mathbb{A}_+\mathring{\mathbf{x}}_1+\frac{1}{2}\mathring{\mathbf{x}}_2'\mathbb{A}_+\mathring{\mathbf{x}}_2-\frac{a_+^2}{2}\sum_{j,k}\mathfrak{s}_{jk}^v(\mathring{x}_{j,1}-\mathring{x}_{j,2})(\mathring{x}_{k,1}-\mathring{x}_{k,2})+R_+^x+R_+^{v,x}\\
&= \frac{1}{2}\mathring{\mathbf{x}}'\mathbb{A}_+^v\mathring{\mathbf{x}}+R_+^x+R_+^{v,x},
\end{aligned}
$$

where

$$
\mathring{\mathbf{x}}:=(\mathring{\mathbf{x}}_1',\mathring{\mathbf{x}}_2')',\qquad\mathbb{A}_+^v:=(1+a_+^2)I_{2W}+a_+^2\mathbb{S}^v,
\tag{8.35}
$$

and recall that $\mathbb{S}^v$ is defined in (5.22). Analogously, for the Type III vicinity, we can write

$$
M(K(\hat{X},V)-K(D_-,I))=\frac{1}{2}\mathring{\mathbf{x}}'\mathbb{A}_-^v\mathring{\mathbf{x}}+R_-^x+R_-^{v,x},
\tag{8.36}
$$

where

$$
\mathbb{A}_-^v:=(1+a_-^2)I_{2W}+a_-^2\mathbb{S}^v,\qquad R_-^{v,x}=O\Big(\frac{\Theta^{\frac{3}{2}}}{\sqrt{M}}\Big).
\tag{8.37}
$$

Consequently, by (8.12) and (8.13) we can write (8.6) for $\kappa=+,-$ as

$$
\begin{aligned}
\mathcal{I}(\Upsilon_+^b,\Upsilon_-^b,\Upsilon_\varkappa^x,\Upsilon_\varkappa^x,\Upsilon_S,\mathbb{I}^{W-1}) &= \exp\Big\{M\big(K(D_\pm,I)-K(D_\varkappa,I)\big)\Big\}\cdot\frac{(-a_\varkappa^2)^W}{(n!)^2}\cdot\frac{M^{W+1}}{8^W\pi^{3W+3}}\\
&\times\int_{\mathbb{L}^{W-1}}\prod_{j=2}^W\frac{\mathrm{d}\theta_j}{2\pi}\int_{\mathbb{I}^{W-1}}\prod_{j=2}^W 2v_j\mathrm{d}v_j\int_{\mathring{\Upsilon}}\prod_{j=1}^W\mathrm{d}\mathring{b}_{j,1}\int_{\mathring{\Upsilon}}\prod_{j=1}^W\mathrm{d}\mathring{b}_{j,2}\int_{\mathring{\Upsilon}}\prod_{j=1}^W\mathrm{d}\mathring{x}_{j,1}\int_{\mathring{\Upsilon}}\prod_{j=1}^W\mathrm{d}\mathring{x}_{j,2}
\end{aligned}
$$

$$\times \int_{\mathring{\Upsilon}_S} \prod_{j=2}^{W} \mathrm{d}\tau_{j,1} \int_{\mathring{\Upsilon}_S} \prod_{j=2}^{W} \mathrm{d}\tau_{j,2} \cdot \exp\Big\{ -\frac{1}{2}\mathring{\mathbf{b}}_1' \mathbb{A}_+ \mathring{\mathbf{b}}_1 - \frac{1}{2}\mathring{\mathbf{b}}_2' \mathbb{A}_- \mathring{\mathbf{b}}_2 - R^b \Big\}$$
$$\times \exp\Big\{ -\frac{1}{2}\mathring{\mathbf{x}}' \mathbb{A}^v_{\varkappa} \mathring{\mathbf{x}} - R^x_{\varkappa} - R^{v,x}_{\varkappa} \Big\} \cdot \exp\Big\{ (a_+ - a_-)^2 \sum_{a=1,2} \boldsymbol{\tau}_a' S^{(1)} \boldsymbol{\tau}_a - R^{t,b} \Big\}$$
$$\times \prod_{j=1}^{W} \exp\Big\{ \mathbf{i} \frac{\mathring{x}_{j,1} + \mathring{x}_{j,2}}{\sqrt{M}} \Big\} \cdot \prod_{j=1}^{W} (x_{j,1} - x_{j,2})^2 (b_{j,1} + b_{j,2})^2 \cdot \mathsf{A}(\hat{X}, \hat{B}, V, T) + O(e^{-\Theta}). \tag{8.38}$$

8.2. **Steepest descent paths in the vicinities.** In order to estimate the integrals (8.32) and (8.38) properly, we need to control various remainder terms in (8.32) and (8.38) to reduce these integrals to Gaussian ones. The final result is collected in Proposition 8.6 at the end of this section. As a preparation, we shall further deform the contours of $\mathring{\mathbf{b}}$-variables and $\mathring{\mathbf{x}}$-variables to the steepest descent paths. We mainly provide the discussion for the $\mathring{\mathbf{b}}$-variables, that for the $\mathring{\mathbf{x}}$-variables is analogous.

For simplicity, in this section, we assume $0 \le E \le \sqrt{2} - \kappa$, the case $-\sqrt{2} + \kappa \le E \le 0$ can be discussed similarly. We introduce the eigendecomposition of S as

$$S = \mathsf{U} \hat{S} \mathsf{U}'.$$

Note that U is an orthogonal matrix thus the entries are all real. Now, we perform the change of coordinate

$$\mathbf{c}_a = (c_{1,a}, \dots, c_{W,a})' := \mathsf{U}' \mathring{\mathbf{b}}_a, \quad a = 1, 2.$$

Obviously, for the differentials, we have

$$\prod_{j=1}^{W} \mathrm{d}\mathring{b}_{j,a} = \prod_{j=1}^{W} \mathrm{d}c_{j,a}, \quad a = 1, 2.$$

In addition, for the domains, it is elementary to see

$$\mathring{\mathbf{b}}_a \in \mathring{\Upsilon} \Longleftrightarrow \mathbf{c}_a \in \mathring{\Upsilon}, \quad a = 1, 2. \tag{8.39}$$

Now, we introduce the notation

$$\gamma_j^+ := \frac{1}{\sqrt{1 + a_+^2 + a_+^2 \lambda_j(S)}}, \qquad \gamma_j^- := \frac{1}{\sqrt{1 + a_-^2 + a_-^2 \lambda_j(S)}},$$

and set the diagonal matrices

$$\mathbb{D}_+ := \mathrm{diag}(\gamma_1^+, \dots, \gamma_W^+), \qquad \mathbb{D}_- := \mathrm{diag}(\gamma_1^-, \dots, \gamma_W^-).$$

By the assumption $0 \le E \le \sqrt{2} - \kappa$ and (1.4), it is not difficult to check

$$|\gamma_j^+| \sim 1, \quad |\gamma_j^-| \sim 1, \qquad \arg \gamma_j^+ \in \big(-\frac{\pi}{8}, 0 \big], \qquad \arg \gamma_j^- \in \big[0, \frac{\pi}{8} \big), \quad \forall\, j = 1, \dots, W. \tag{8.40}$$

With the notation introduced above, we have

$$\mathring{\mathbf{b}}_1' \mathbb{A}_+ \mathring{\mathbf{b}}_1 = \mathbf{c}_1' \mathbb{D}_+^{-2} \mathbf{c}_1, \quad \mathring{\mathbf{b}}_2' \mathbb{A}_- \mathring{\mathbf{b}}_2 = \mathbf{c}_2' \mathbb{D}_-^{-2} \mathbf{c}_2.$$

To simplify the following discussion, we enlarge the domain of the $\mathbf{c}$-variables to

$$\mathbf{c}_a \in \Upsilon_\infty \equiv \Upsilon_\infty(\varepsilon) := [-\Theta^{\frac{1}{2}}, \Theta^{\frac{1}{2}}]^W, \quad a = 1, 2.$$

Obviously, $\mathring{\Upsilon} \subset \Upsilon_\infty$. It is easy to check that (7.26) also holds when $\mathbf{c}_a \in \Upsilon_\infty \setminus \mathring{\Upsilon}$ for either $a = 1$ or 2, according to (8.39), thus such a modification of the domain will only produce an error term of order $O(\exp\{-\Theta\})$ in the integral (8.32), by using (8.7).

Now we do the scaling

$$\mathbf{c}_1 \to \mathbb{D}_+ \mathbf{c}_1, \qquad \mathbf{c}_2 \to \mathbb{D}_- \mathbf{c}_2.$$

Consequently, we have

$$\mathring{\mathbf{b}}_1 = \mathsf{U}\mathbb{D}_+\mathbf{c}_1, \qquad \mathring{\mathbf{b}}_2 = \mathsf{U}\mathbb{D}_-\mathbf{c}_2, \tag{8.41}$$

thus

$$\mathring{\mathbf{b}}_1'\mathbb{A}_+\mathring{\mathbf{b}}_1 = \sum c_{j,1}^2, \quad \mathring{\mathbf{b}}_2'\mathbb{A}_-\mathring{\mathbf{b}}_2 = \sum c_{j,2}^2. \tag{8.42}$$

Accordingly, we should adjust the change of differentials as

$$\prod_{j=1}^W \mathrm{d}\mathring{b}_{j,1} \to \det\mathbb{D}_+ \cdot \prod_{j=1}^W \mathrm{d}c_{j,1}, \quad \prod_{j=1}^W \mathrm{d}\mathring{b}_{j,2} \to \det\mathbb{D}_- \cdot \prod_{j=1}^W \mathrm{d}c_{j,2}.$$

In addition, the domain of $\mathbf{c}_1$ should be changed from Υ_∞ to $\prod_{j=1}^W \mathbb{J}_j^+$, where

$$\mathbb{J}_j^+ := (\gamma_j^+)^{-1}[-\Theta^{\frac12}, \Theta^{\frac12}],$$

and that of $\mathbf{c}_2$ should be changed from Υ_∞ to $\prod_{j=1}^W \mathbb{J}_j^-$, where

$$\mathbb{J}_j^- := (\gamma_j^-)^{-1}[-\Theta^{\frac12}, \Theta^{\frac12}].$$

By the fact $\det\mathbb{D}_+\mathbb{D}_- = 1/\sqrt{\det\mathbb{A}_+\mathbb{A}_-}$, we can write (8.32) as

$$\begin{aligned}
2^W\mathcal{I}(\Upsilon_+^b, \Upsilon_-^b, \Upsilon_+^x, \Upsilon_-^x, \Upsilon_S, \Upsilon_S) &= \frac{M^2}{(n!)^2 4^W \pi^{4W+2}} \cdot \frac{1}{\sqrt{\det\mathbb{A}_+\mathbb{A}_-}} \\
&\times \int_{\prod_{j=1}^W \mathbb{J}_j^+} \prod_{j=1}^W \mathrm{d}c_{j,1} \int_{\prod_{j=1}^W \mathbb{J}_j^-} \prod_{j=1}^W \mathrm{d}c_{j,2} \int_{\mathring{\Upsilon}} \prod_{j=1}^W \mathrm{d}\mathring{x}_{j,1} \int_{\mathring{\Upsilon}} \prod_{j=1}^W \mathrm{d}\mathring{x}_{j,2} \\
&\times \int_{\mathring{\Upsilon}_S} \prod_{j=2}^W \mathrm{d}\tau_{j,1} \int_{\mathring{\Upsilon}_S} \prod_{j=2}^W \mathrm{d}\tau_{j,2} \int_{\mathring{\Upsilon}_S} \prod_{j=2}^W \mathrm{d}\upsilon_{j,1} \int_{\mathring{\Upsilon}_S} \prod_{j=2}^W \mathrm{d}\upsilon_{j,2} \prod_{j=1}^W \exp\Big\{\mathbf{i}\frac{\mathring{x}_{j,1}+\mathring{x}_{j,2}}{\sqrt{M}}\Big\} \\
&\times \exp\Big\{-\frac12\sum c_{j,1}^2 - \frac12\sum c_{j,2}^2 - R^b\Big\} \cdot \exp\Big\{-\frac12\mathring{\mathbf{x}}_1'\mathbb{A}_+\mathring{\mathbf{x}}_1 - \frac12\mathring{\mathbf{x}}_2'\mathbb{A}_-\mathring{\mathbf{x}}_2 - R_\pm^x\Big\} \\
&\times \exp\Big\{(a_+-a_-)^2\sum_{a=1,2}\boldsymbol{\tau}_a'S^{(1)}\boldsymbol{\tau}_a - R^{t,b}\Big\} \cdot \exp\Big\{(a_+-a_-)^2\sum_{a=1,2}\boldsymbol{\upsilon}_a'S^{(1)}\boldsymbol{\upsilon}_a - R_\pm^{v,x}\Big\} \\
&\times \prod_{j=1}^W (x_{j,1}-x_{j,2})^2(b_{j,1}+b_{j,2})^2 \cdot \mathsf{A}(\hat{X}, \hat{B}, V, T) + O(e^{-\Theta}).
\end{aligned} \tag{8.43}$$

In (8.43), all the $\mathbf{b}$ and $\mathring{\mathbf{b}}$-variables in the integrand should be regarded as the functions of the $\mathbf{c}$-variables, see (8.41).

Now, we consider the integrand as a function of $\mathbf{c}$-variables on the disks, namely,

$$c_{j,1} \in \mathbb{O}_j^+ := \big\{z \in \mathbb{C} : |z| \le \Theta^{\frac12}|\gamma_j^+|^{-1}\big\}, \quad c_{j,2} \in \mathbb{O}_j^- := \big\{z \in \mathbb{C} : |z| \le \Theta^{\frac12}|\gamma_j^-|^{-1}\big\}.$$

For $\mathbf{c}_1 \in \prod_{j=1}^W \mathbb{O}_j^+$ and $\mathbf{c}_2 \in \prod_{j=1}^W \mathbb{O}_j^-$, by (8.40) and (8.41) we have

$$||\mathring{\mathbf{b}}_1||_\infty, ||\mathring{\mathbf{b}}_2||_\infty \le O(\Theta). \tag{8.44}$$

Here we used the elementary fact $||U\mathbf{a}||_\infty \le \sqrt{W}||\mathbf{a}||_\infty$ for any $\mathbf{a} \in \mathbb{C}^W$ and and unitary matrix U. Then, we deform the contour of $c_{j,1}$ from $\mathbb{J}_j^+$ to

$$(-\Sigma_j^+) \cup \mathbb{L}_j^+ \cup \Sigma_j^+$$

for each $j = 1, \dots, W$, where

$$\mathbb{L}_j^+ := \mathbb{R} \cap \mathbb{O}_j^+, \qquad \Sigma_j^+ = \big\{z \in \partial\mathbb{O}_j^+ : 0 \le \arg z \le -\arg\gamma_j^+\big\}.$$

It is not difficult to see that

$$\mathsf{Re} c_{j,1}^2 \geq \Theta, \quad \text{for} \quad c_{j,1} \in (-\Sigma_j^+) \cup \Sigma_j^+,$$

by using (8.40). Consequently, by (8.42), we have

$$\Big| \exp\Big\{ -\frac{1}{2} \mathring{\mathbf{b}}_1' \mathbb{A}_+ \mathring{\mathbf{b}}_1 \Big\} \Big| = \Big| \exp\Big\{ -\frac{1}{2} \sum_{j=1}^W c_{j,1}^2 \Big\} \Big| \leq O(e^{-\Theta}).$$

Then using (8.7), we can get rid of the integral over Σ_j^+ and $-\Sigma_j^+$, analogously to the discussion in Section 7. Similarly, we can perform the same argument for $\mathbf{c}_2$. Consequently, we can restrict the integral in (8.43) to the domain

$$\mathbf{c}_1 \in \prod_{j=1}^W \mathbb{L}_j^+, \qquad \mathbf{c}_2 \in \prod_{j=1}^W \mathbb{L}_j^-.$$

So we can assume that $\prod_{j=1}^W \mathbb{J}_j^+$ and $\prod_{j=1}^W \mathbb{J}_j^-$ are replaced with $\prod_{j=1}^W \mathbb{L}_j^+$ and $\prod_{j=1}^W \mathbb{L}_j^-$ respectively in (8.43).

By (8.15), (8.44) and the fact $||\mathbf{a}||_3^3 \leq ||\mathbf{a}||_\infty ||\mathbf{a}||_2^2$ for any vector $\mathbf{a}$, we see that

$$|R^b| \leq C \frac{||\mathring{\mathbf{b}}_1||_\infty + ||\mathring{\mathbf{b}}_2||_\infty}{\sqrt{M}} \big(||\mathring{\mathbf{b}}_1||_2^2 + ||\mathring{\mathbf{b}}_2||_2^2 \big) \leq \frac{\Theta}{\sqrt{M}} \big(||\mathbf{c}_1||_2^2 + ||\mathbf{c}_2||_2^2 \big) \tag{8.45}$$

for some positive constant C, where in the last step we also used the fact that $||\mathbf{b}_a||_2 = O(||\mathbf{c}_a||_2)$ for $a = 1, 2$, which is implied by (8.41) and (8.40). Consequently, we have

$$\exp\Big\{ -\frac{1}{2}||\mathbf{c}_1||_2^2 - \frac{1}{2}||\mathbf{c}_2||_2^2 - R^b \Big\} = \exp\Big\{ -\big(\frac{1}{2} + o(1)\big)||\mathbf{c}_1||_2^2 - \big(\frac{1}{2} + o(1)\big)||\mathbf{c}_2||_2^2 \Big\}.$$

This allows us to take a step further to truncate $\mathbf{c}_1$ and $\mathbf{c}_2$ according to their 2-norm, namely

$$\mathbf{c}_1, \mathbf{c}_2 \in \mathring{\Upsilon}. \tag{8.46}$$

Similarly to the discussion in the proof of Lemma 7.7, such a truncation will only produce an error of order $\exp\{-\Theta\}$ to the integral, by using (8.7).

Now, analogously to (8.41), we can change $\mathring{\mathbf{x}}$-variables to $\mathbf{d}$-variables, defined by

$$\mathbf{d}_1 = (d_{1,1}, \ldots, d_{W,1}) := \mathbb{D}_+^{-1} \mathsf{U}' \mathring{\mathbf{x}}_1, \qquad \mathbf{d}_2 = (d_{1,2}, \ldots, d_{W,2}) := \mathbb{D}_-^{-1} \mathsf{U}' \mathring{\mathbf{x}}_2.$$

Thus accordingly, we change the differentials

$$\prod_{j=1}^W \mathrm{d}\mathring{x}_{j,1} \to \det \mathbb{D}_+ \cdot \prod_{j=1}^W \mathrm{d}d_{j,1}, \quad \prod_{j=1}^W \mathrm{d}\mathring{x}_{j,2} \to \det \mathbb{D}_- \cdot \prod_{j=1}^W \mathrm{d}d_{j,2}.$$

In addition, like (8.46), we deform the domain to

$$\mathbf{d}_1, \mathbf{d}_2 \in \mathring{\Upsilon}.$$

Finally, from (8.43), we arrive at the representation

$$\begin{aligned}
&2^W \mathcal{I}(\Upsilon_+^b, \Upsilon_-^b, \Upsilon_+^x, \Upsilon_-^x, \Upsilon_S, \Upsilon_S) \\
&= \frac{M^2}{(n!)^2 4^W \pi^{4W+2}} \cdot \frac{1}{\det \mathbb{A}_+ \mathbb{A}_-} \cdot \int_{\mathring{\Upsilon}} \prod_{j=1}^W \mathrm{d}c_{j,1} \int_{\mathring{\Upsilon}} \prod_{j=1}^W \mathrm{d}c_{j,2} \int_{\mathring{\Upsilon}} \prod_{j=1}^W \mathrm{d}d_{j,1} \int_{\mathring{\Upsilon}} \prod_{j=1}^W \mathrm{d}d_{j,2} \\
&\times \int_{\mathring{\Upsilon}_S} \prod_{j=2}^W \mathrm{d}\tau_{j,1} \int_{\mathring{\Upsilon}_S} \prod_{j=2}^W \mathrm{d}\tau_{j,2} \int_{\mathring{\Upsilon}_S} \prod_{j=2}^W \mathrm{d}\upsilon_{j,1} \int_{\mathring{\Upsilon}_S} \prod_{j=2}^W \mathrm{d}\upsilon_{j,2} \prod_{j=1}^W \exp\Big\{ \mathbf{i} \frac{\mathring{x}_{j,1} + \mathring{x}_{j,2}}{\sqrt{M}} \Big\} \\
&\times \exp\Big\{ -\frac{1}{2}||\mathbf{c}_1||_2^2 - \frac{1}{2}||\mathbf{c}_2||_2^2 - R^b \Big\} \cdot \exp\Big\{ -\frac{1}{2}||\mathbf{d}_1||_2^2 - \frac{1}{2}||\mathbf{d}_2||_2^2 - R_\pm^x \Big\}
\end{aligned}$$

$$\times \exp\Big\{(a_+-a_-)^2\sum_{a=1,2}\boldsymbol{\tau}_a' S^{(1)}\boldsymbol{\tau}_a - R^{t,b}\Big\}\cdot\exp\Big\{(a_+-a_-)^2\sum_{a=1,2}\boldsymbol{\upsilon}_a' S^{(1)}\boldsymbol{\upsilon}_a - R_\pm^{v,x}\Big\}$$
$$\times\prod_{j=1}^W (x_{j,1}-x_{j,2})^2(b_{j,1}+b_{j,2})^2\cdot \mathsf{A}(\hat X,\hat B,V,T)+O(e^{-\Theta}), \tag{8.47}$$

in which $\mathbf{x}$ and $\mathring{\mathbf{x}}$-variables should be regarded as functions of the $\mathbf{d}$-variables, as well, $\mathbf{b}$ and $\mathring{\mathbf{b}}$-variables should be regarded as functions of the $\mathbf{c}$-variables.

Now, in the Type II and III vicinities, we only do the change of coordinates for the $\mathring{\mathbf{b}}$-variables, which is enough for our purpose. Consequently, we have

$$\mathcal{I}\big(\Upsilon_+^b,\Upsilon_-^b,\Upsilon_\varkappa^x,\Upsilon_\varkappa^x,\Upsilon_S,\mathbb{I}^{W-1}\big)=\exp\big\{M(K(D_\pm,I)-K(D_\varkappa,I))\big\}\cdot\frac{(-a_\varkappa^2)^W}{(n!)^2}\cdot\frac{M^{W+1}}{8^W\pi^{3W+3}}$$
$$\times\frac{1}{\sqrt{\det\mathbb{A}_+\mathbb{A}_-}}\cdot\int_{\mathbb{L}^{W-1}}\prod_{j=2}^W\frac{\mathrm{d}\theta_j}{2\pi}\int_{\mathbb{I}^{W-1}}\prod_{j=2}^W 2v_j\mathrm{d}v_j\int_{\mathring\Upsilon}\prod_{j=1}^W\mathrm{d}c_{j,1}\int_{\mathring\Upsilon}\prod_{j=1}^W\mathrm{d}c_{j,2}$$
$$\times\int_{\mathring\Upsilon}\prod_{j=1}^W\mathrm{d}\mathring x_{j,1}\int_{\mathring\Upsilon}\prod_{j=1}^W\mathrm{d}\mathring x_{j,2}\int_{\mathring\Upsilon_S}\prod_{j=2}^W\mathrm{d}\tau_{j,1}\int_{\mathring\Upsilon_S}\prod_{j=2}^W\mathrm{d}\tau_{j,2}\prod_{j=1}^W\exp\Big\{\mathbf{i}\frac{\mathring x_{j,1}+\mathring x_{j,2}}{\sqrt{M}}\Big\}$$
$$\times\exp\Big\{-\frac12||\mathbf{d}_1||_2^2-\frac12||\mathbf{d}_2||_2^2-R^b\Big\}\cdot\exp\Big\{(a_+-a_-)^2\sum_{a=1,2}\boldsymbol{\tau}_a' S^{(1)}\boldsymbol{\tau}_a-R^{t,b}\Big\}$$
$$\times\exp\Big\{-\frac12\mathring{\mathbf{x}}'\mathbb{A}_\varkappa^v\mathring{\mathbf{x}}-R_\varkappa^x-R_\varkappa^{v,x}\Big\}\prod_{j=1}^W(x_{j,1}-x_{j,2})^2(b_{j,1}+b_{j,2})^2\cdot\mathsf{A}(\hat X,\hat B,V,T)+O(e^{-\Theta}). \tag{8.48}$$

By (8.40) and (8.41), it is easy to see

$$\mathbf{c}_1,\mathbf{c}_2\in\mathring\Upsilon\Longrightarrow\mathring{\mathbf{b}}_1,\mathring{\mathbf{b}}_2\in\mathring\Upsilon \tag{8.49}$$

Similarly, we have

$$\mathbf{d}_1,\mathbf{d}_2\in\mathring\Upsilon\Longrightarrow\mathring{\mathbf{x}}_1,\mathring{\mathbf{x}}_2\in\mathring\Upsilon. \tag{8.50}$$

We keep the terminology "Type I', II and III vicinities" for the slightly modified domains defined in terms of $\mathbf{c}$, $\mathbf{d}$, $\boldsymbol{\tau}$ and $\boldsymbol{\upsilon}$-variables. More specifically, we redefine the vicinities as follows.

Definition 8.5. *We slightly modify Definition 8.1 as follows.*

- *Type I' vicinity:* $\mathbf{c}_1,\mathbf{c}_2,\mathbf{d}_1,\mathbf{d}_2\in\mathring\Upsilon$, $\boldsymbol{\tau}_1,\boldsymbol{\tau}_2,\boldsymbol{\upsilon}_1,\boldsymbol{\upsilon}_2\in\mathring\Upsilon_S$.
- *Type II vicinity:* $\mathbf{c}_1,\mathbf{c}_2,\mathring{\mathbf{x}}_1,\mathring{\mathbf{x}}_2\in\mathring\Upsilon$, $\boldsymbol{\tau}_1,\boldsymbol{\tau}_2\in\mathring\Upsilon_S$, $V_j\in\mathring U(2)$ *for all* $j=2,\dots,W$, *where* $\mathring{\mathbf{x}}$*-variables are defined in (8.8) with* $\varkappa=+$.
- *Type III vicinity:* $\mathbf{c}_1,\mathbf{c}_2,\mathring{\mathbf{x}}_1,\mathring{\mathbf{x}}_2\in\mathring\Upsilon$, $\boldsymbol{\tau}_1,\boldsymbol{\tau}_2\in\mathring\Upsilon_S$, $V_j\in\mathring U(2)$ *for all* $j=2,\dots,W$, *where* $\mathring{\mathbf{x}}$*-variables are defined in (8.8) with* $\varkappa=-$.

Now, recall the remainder terms R^b, $R_\pm^x$, R_+^x and R_-^x in Lemma 8.2, $R^{t,b}$ and $R_\pm^{v,x}$ in (8.23), $R_+^{v,x}$ in (8.33) and $R_-^{v,x}$ in (8.36). In light of (8.49) and (8.50), the bounds on these remainder terms are the same as those obtained in Section 8.1. For the convenience of the reader, we collect them as the following proposition.

Proposition 8.6. *Under Assumptions 1.1 and 1.13, we have the following estimate, in the vicinities.*

$$(i):\quad R^{t,b}=O\Big(\frac{\Theta^4}{M}\Big),\quad R_\pm^{v,x}=O\Big(\frac{\Theta^4}{M}\Big),$$

$$(ii): \quad R^b = O\Big(\frac{\Theta^2}{\sqrt{M}}\Big), \quad R^x = O\Big(\frac{\Theta^2}{\sqrt{M}}\Big), \quad R^x_+ = O\Big(\frac{\Theta^{\frac{3}{2}}}{\sqrt{M}}\Big), \quad R^x_- = O\Big(\frac{\Theta^{\frac{3}{2}}}{\sqrt{M}}\Big),$$
$$(iii): \quad R^{v,x}_+ = O\Big(\frac{\Theta^{\frac{3}{2}}}{\sqrt{M}}\Big), \quad R^{v,x}_- = O\Big(\frac{\Theta^{\frac{3}{2}}}{\sqrt{M}}\Big).$$

Proof. Note that, (i) can be obtained from (8.26), and (ii) follows from Lemma 8.2, and (iii) is implied by (8.34) and (8.37). Hence, we completed the proof. □

Analogously, in the vicinities, $||\mathring{\mathbf{b}}_1||_2^2$, $||\mathring{\mathbf{b}}_2||_2^2$, $||\mathring{\mathbf{x}}_1||_2^2$, $||\mathring{\mathbf{x}}_2||_2^2$, $||\mathring{\mathbf{t}}||_\infty$ and $||\mathring{\mathbf{v}}||_\infty$ are still bounded by Θ.

9. Integral over the Type I vicinities

With (8.47), we estimate the integral over the Type I vicinity in this section. At first, in the Type I' vicinity, we have $||\mathring{\mathbf{x}}_a||_\infty = O(\Theta^{\frac{1}{2}})$ and $||\mathring{\mathbf{b}}_a||_\infty = O(\Theta^{\frac{1}{2}})$ for $a = 1, 2$. Consequently, according to the parametrization in (8.2), we have

$$x_{j,1} - x_{j,2} = a_+ - a_- + O\Big(\frac{\Theta^{\frac{1}{2}}}{\sqrt{M}}\Big), \qquad b_{j,1} + b_{j,2} = a_+ - a_- + O\Big(\frac{\Theta^{\frac{1}{2}}}{\sqrt{M}}\Big), \tag{9.1}$$

which implies

$$\prod_{j=1}^{W}(x_{j,1} - x_{j,2})^2(b_{j,1} + b_{j,2})^2 = (a_+ - a_-)^{4W}\Big(1 + O\Big(\frac{\Theta^{\frac{3}{2}}}{\sqrt{M}}\Big)\Big). \tag{9.2}$$

Hence, what remains is to estimate the function $\mathsf{A}(\hat{X}, \hat{B}, V, T)$. We have the following lemma.

Lemma 9.1. *Suppose that the assumptions in Theorem 1.14 hold. In the Type I' vicinity, for any given positive integer n, there is $N_0 = N_0(n)$, such that for all $N \geq N_0$ we have*

$$|\mathsf{A}(\hat{X}, \hat{B}, V, T)| \leq \frac{\Theta^2 W^{C_0}}{M(N\eta)^{n+\ell}} \cdot |\det \mathbb{A}_+|^2 \cdot \det(S^{(1)})^2.$$

for some positive constant C_0 and some integer $\ell = O(1)$, both of which are independent of n.

With (8.47), (9.2) and Lemma 9.1, we can prove Lemma 5.8.

Proof of Lemma 5.8. Using (8.47), (9.2), Lemma 9.1, Proposition 8.6 with (5.38), the fact $\det \mathbb{A}_+ = \overline{\det \mathbb{A}_-}$ and the trivial estimate

$$M\Theta^2 W^{C_0}\frac{1}{(N\eta)^\ell} \leq N^{C_0}$$

for sufficiently large constant C_0, we have

$$\begin{aligned}
2^W|\mathcal{I}(\Upsilon^b_+, \Upsilon^b_-, \Upsilon^x_+, \Upsilon^x_-, \Upsilon_S, \Upsilon_S)| &\leq \frac{N^{C_0}}{(N\eta)^n} \cdot \frac{1}{(2\pi^2)^{2W}} \cdot \det(S^{(1)})^2 \cdot (a_+ - a_-)^{4W} \\
&\times \int_{\mathring{\Upsilon}} \prod_{j=1}^{W} \mathrm{d}c_{j,1} \int_{\mathring{\Upsilon}} \prod_{j=1}^{W} \mathrm{d}c_{j,2} \int_{\mathring{\Upsilon}} \prod_{j=1}^{W} \mathrm{d}d_{j,1} \int_{\mathring{\Upsilon}} \prod_{j=1}^{W} \mathrm{d}d_{j,2} \int_{\mathring{\Upsilon}_S} \prod_{j=2}^{W} \mathrm{d}\tau_{j,1} \int_{\mathring{\Upsilon}_S} \prod_{j=2}^{W} \mathrm{d}\tau_{j,2} \\
&\times \int_{\mathring{\Upsilon}_S} \prod_{j=2}^{W} \mathrm{d}\upsilon_{j,1} \int_{\mathring{\Upsilon}_S} \prod_{j=2}^{W} \mathrm{d}\upsilon_{j,2} \ \exp\Big\{-\frac{1}{2}\big(||\mathbf{c}_1||_2^2 + ||\mathbf{c}_2||_2^2 + ||\mathbf{d}_1||_2^2 + ||\mathbf{d}_2||_2^2\big)\Big\} \\
&\times \exp\Big\{(a_+ - a_-)^2\Big(\boldsymbol{\tau}_1' S^{(1)}\boldsymbol{\tau}_1 + \boldsymbol{\tau}_1' S^{(1)}\boldsymbol{\tau}_2 + \boldsymbol{\upsilon}_1' S^{(1)}\boldsymbol{\upsilon}_1 + \boldsymbol{\upsilon}_1' S^{(1)}\boldsymbol{\upsilon}_2\Big)\Big\} + O(e^{-\Theta}).
\end{aligned}$$

Then, by elementary Gaussian integral we obtain (5.40). Hence, we completed the proof of Lemma 5.8. □

The remaining part of this section will be dedicated to the proof of Lemma 9.1. Recall the definitions of the functions $\mathsf{A}(\cdot)$, $\mathsf{Q}(\cdot)$, $\mathsf{P}(\cdot)$ and $\mathsf{F}(\cdot)$ in (3.31), (4.3), (4.4) and (4.5). Using the strategy in Section 6 again, we ignore the irrelevant factor $\mathsf{Q}(\cdot)$ at the beginning. Hence, we bound $\mathsf{P}(\cdot)$ and $\mathsf{F}(\cdot)$ at first, and modify the bounding procedure slightly to take $\mathsf{Q}(\cdot)$ into account in the end, resulting a proof of Lemma 9.1.

9.1. $\mathsf{P}(\hat{X}, \hat{B}, V, T)$ **in the Type I' vicinity.** As mentioned above, we should always regard $\mathbf{b}$ or $\mathring{\mathbf{b}}$-variables as functions of $\mathbf{c}$-variables, regard $\mathbf{x}$ or $\mathring{\mathbf{x}}$-variables as functions of $\mathbf{d}$-variables. Our aim, in this section, is to prove the following lemma.

Lemma 9.2. *Suppose that the assumptions in Theorem 1.14 hold. In the Type I' vicinity, we have*

$$\mathsf{P}(\hat{X}, \hat{B}, V, T) \le \frac{W^{2+\gamma}\Theta^2}{M} |\det \mathbb{A}_+|^2 \det(S^{(1)})^2. \tag{9.3}$$

Before commencing the formal proof, we introduce more notation below. In the sequel, we will use the notation

$$\mathring{\kappa}_j \equiv \mathring{\kappa}_j(\hat{X}, \hat{B}, V, T) := |\mathring{x}_{j,1}| + |\mathring{x}_{j,2}| + |\mathring{b}_{j,1}| + |\mathring{b}_{j,2}| + |\mathring{v}_j| + |\mathring{t}_j| = O(\Theta), \tag{9.4}$$

where the bound holds in the Type I' vicinity, according to the facts $||\mathring{\mathbf{x}}_a||_\infty = O(\Theta^{\frac{1}{2}})$, $||\mathring{\mathbf{b}}_a||_\infty = O(\Theta^{\frac{1}{2}})$ for $a = 1, 2$, $||\mathring{\mathbf{t}}||_\infty = O(\Theta)$ and $||\mathring{\mathbf{v}}||_\infty = O(\Theta)$.

Recalling (6.17) with $\boldsymbol{\varpi}_j$ defined in (6.13) and $\hat{\boldsymbol{\varpi}}_j$ in (6.15). Now, we write

$$\begin{aligned} \boldsymbol{\varpi}_j &= \exp\Big\{ - M \log\det\big(1 + M^{-1} V_j^* \hat{X}_j^{-1} V_j \Omega_j T_j^{-1} \hat{B}_j^{-1} T_j \Xi_j\big)\Big\} \\ &=: \exp\Big\{ - Tr V_j^* \hat{X}_j^{-1} V_j \Omega_j T_j^{-1} \hat{B}_j^{-1} T_j \Xi_j \Big\} \exp\Big\{ \sum_{\ell=2}^{4} \frac{(-1)^{\ell-1}}{\ell M^{\ell-1}} \Delta_{\ell,j} \Big\}, \end{aligned} \tag{9.5}$$

where

$$\Delta_{\ell,j} := Tr\big(V_j^* \hat{X}_j^{-1} V_j \Omega_j T_j^{-1} \hat{B}_j^{-1} T_j \Xi_j\big)^\ell. \tag{9.6}$$

The second step of (9.5) follows from the Taylor expansion of the logarithmic function. The expansion terminates at the 4th order term since $\Delta_{\ell,j}$ is a homogeneous polynomial of Ω_j and Ξ_j-variables with degree 2ℓ, regarding all the complex variables as fixed parameters. Now, we expand the first factor of (9.5) around the Type I' saddle point, namely

$$\exp\Big\{ - Tr V_j^* \hat{X}_j^{-1} V_j \Omega_j T_j^{-1} \hat{B}_j^{-1} T_j \Xi_j \Big\} =: \exp\Big\{ - Tr D_\pm^{-1} \Omega_j D_\pm^{-1} \Xi_j \Big\} \exp\Big\{ - \frac{1}{\sqrt{M}} \Delta_j \Big\}. \tag{9.7}$$

We take (9.7) as the definition of Δ_j, which is of the form

$$\Delta_j = \sum_{\alpha,\beta=1}^{4} \mathring{\mathfrak{p}}_{j,\alpha,\beta} \cdot \omega_{j,\alpha} \xi_{j,\beta}$$

for some function $\mathring{\mathfrak{p}}_{j,\alpha,\beta}$ of $\mathring{\mathbf{x}}$, $\mathring{\mathbf{b}}$, $\mathring{\mathbf{v}}$ and $\mathring{\mathbf{t}}$-variables, satisfying

$$\mathring{\mathfrak{p}}_{j,\alpha,\beta} = O(\mathring{\kappa}_j), \qquad \forall\, \alpha, \beta = 1, \dots, 4. \tag{9.8}$$

One can check (9.8) easily by using (8.8)-(8.10). Analogously, we can also write

$$\Delta_{\ell,j} = \sum_{\substack{\alpha_1,\dots,\alpha_\ell,\\ \beta_1,\dots,\beta_\ell=1}}^{4} \mathring{\mathfrak{p}}_{\ell,j,\boldsymbol{\alpha},\boldsymbol{\beta}} \prod_{i=1}^{\ell} \omega_{j,\alpha_i} \xi_{j,\beta_i}, \qquad \boldsymbol{\alpha} := (\alpha_1, \dots, \alpha_\ell), \quad \boldsymbol{\beta} := (\beta_1, \dots, \beta_\ell), \tag{9.9}$$

where

$$\mathring{\mathfrak{p}}_{\ell,j,\boldsymbol{\alpha},\boldsymbol{\beta}} = O(1), \qquad \forall\, \ell = 2, \dots, 4; \alpha_1, \dots, \alpha_\ell, \beta_1, \dots, \beta_\ell = 1, \dots, 4. \tag{9.10}$$

The bound on $\mathring{\mathfrak{p}}_{\ell,j,\boldsymbol{\alpha},\boldsymbol{\beta}}$ in (9.10) follows from the fact that all the V_j, $\hat{X}_j^{-1}$, T_j, T_j^{-1} and $\hat{B}_j^{-1}$-entries are bounded in the Type I' vicinity, uniformly in j. Consequently, we can write for $j \neq p, q$

$$\exp\Big\{ -\frac{1}{\sqrt{M}}\Delta_j + \sum_{\ell=2}^{4}\frac{(-1)^{\ell-1}}{\ell M^{\ell-1}}\Delta_{\ell,j}\Big\} = 1 + \sum_{\ell=1}^{4} M^{-\frac{\ell}{2}} \sum_{\substack{\alpha_1,\ldots,\alpha_\ell,\\ \beta_1,\ldots,\beta_\ell=1}}^{4} \mathring{\mathfrak{q}}_{\ell,j,\boldsymbol{\alpha},\boldsymbol{\beta}} \prod_{i=1}^{\ell} \omega_{j,\alpha_i}\xi_{j,\beta_i}. \tag{9.11}$$

In a similar manner, we can also write for $k = p, q$,

$$\exp\Big\{ -\frac{1}{\sqrt{M}}\Delta_k + \sum_{\ell=2}^{4}\frac{(-1)^{\ell-1}}{\ell M^{\ell-1}}\Delta_{\ell,k}\Big\}\hat{\varpi}_k = \hat{\mathfrak{p}}_0(\cdot)\Big(1 + \sum_{\ell=1}^{4} M^{-\frac{\ell}{2}} \sum_{\substack{\alpha_1,\ldots,\alpha_\ell,\\ \beta_1,\ldots,\beta_\ell=1}}^{4} \mathring{\mathfrak{q}}_{\ell,k,\boldsymbol{\alpha},\boldsymbol{\beta}} \prod_{i=1}^{\ell} \omega_{k,\alpha_i}\xi_{k,\beta_i}\Big), \tag{9.12}$$

where $\hat{\mathfrak{p}}_0(\cdot) = \det \hat{X}_k / \det \hat{B}_k$, which is introduced in (6.15), and $\mathring{\mathfrak{q}}_{\ell,j,\boldsymbol{\alpha},\boldsymbol{\beta}}$ is some function of $\hat{X}$, $\hat{B}$, V and T-variables, satisfying the bound

$$\mathring{\mathfrak{q}}_{\ell,j,\boldsymbol{\alpha},\boldsymbol{\beta}} = O((1+\mathring{\kappa}_j)^{\ell}), \quad \forall\, \ell = 1,\ldots,4, \quad j = 1,\ldots,W. \tag{9.13}$$

Obviously, we have $\hat{\mathfrak{p}}_0(\cdot) = O(1)$ in Type I' vicinity.

Now, in order to distinguish $\ell, \boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ for different j, we index them as $\ell_j, \boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_j$, where

$$\boldsymbol{\alpha}_j \equiv \boldsymbol{\alpha}_j(\ell_j) := (\alpha_{j,1},\ldots,\alpha_{j,\ell_j}), \quad \boldsymbol{\beta}_j \equiv \boldsymbol{\beta}_j(\ell_j) := (\beta_{j,1},\ldots,\beta_{j,\ell_j}).$$

In addition, we introduce the vector

$$\vec{\ell} := (\ell_1,\ldots,\ell_W), \quad \vec{\boldsymbol{\alpha}} \equiv \vec{\boldsymbol{\alpha}}(\vec{\ell}) := (\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_W), \quad \vec{\boldsymbol{\beta}} \equiv \vec{\boldsymbol{\beta}}(\vec{\ell}) := (\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_W).$$

Let $||\vec{\ell}||_1 = \sum_{j=1}^{W}\ell_j$ be the 1-norm of $\vec{\ell}$. Note that $\vec{\boldsymbol{\alpha}}$ and $\vec{\boldsymbol{\beta}}$ are $||\vec{\ell}||_1$-dimensional. With these notations, using (6.17), (9.5), (9.7), (9.11) and (9.12) we have the representation

$$\begin{aligned}\mathcal{P}(\Omega,\Xi,\hat{X},\hat{B},V,T) = \hat{\mathfrak{p}}_0(\hat{X}_p,\hat{B}_p)\hat{\mathfrak{p}}_0(\hat{X}_q,\hat{B}_q)\cdot\exp\Big\{-\sum_{j,k}\tilde{\mathfrak{s}}_{jk}Tr\Omega_j\Xi_k - \sum_{j=1}^{W}TrD_{\pm}^{-1}\Omega_jD_{\pm}^{-1}\Xi_j\Big\}\\ \times\Big(1+\sum_{\substack{\vec{\ell}\in[\![0,4]\!]^W,\\ \text{s.t.}||\vec{\ell}||_1\geq 1}} M^{-\frac{||\vec{\ell}||_1}{2}} \sum_{\vec{\boldsymbol{\alpha}},\vec{\boldsymbol{\beta}}\in[\![1,4]\!]^{||\vec{\ell}||_1}} \prod_{j=1}^{W}\mathring{\mathfrak{q}}_{\ell_j,j,\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j}\cdot\prod_{j=1}^{W}\prod_{i=1}^{\ell_j}\omega_{j,\alpha_{j,i}}\xi_{j,\beta_{j,i}}\Big),\end{aligned} \tag{9.14}$$

where we made the convention

$$\mathring{\mathfrak{q}}_{0,j,\emptyset,\emptyset} = 1, \quad \prod_{i=1}^{0}\omega_{j,\alpha_{j,i}}\xi_{j,\beta_{j,i}} = 1, \quad \forall\, j = 1,\ldots,W. \tag{9.15}$$

According to (9.13) and (9.15), we have

$$\prod_{j=1}^{W}|\mathring{\mathfrak{q}}_{\ell_j,j,\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j}| \leq e^{O(||\vec{\ell}||_1)}\prod_{j=1}^{W}(1+\mathring{\kappa}_j)^{\ell_j}. \tag{9.16}$$

In addition, we can decompose the sum

$$\sum_{\substack{\vec{\ell}\in[\![0,4]\!]^W,\\ \text{s. t.}||\vec{\ell}||_1\geq 1}} = \sum_{\mathfrak{m}=1}^{4W}\sum_{\substack{\vec{\ell}\in[\![0,4]\!]^W,\\ \text{s. t.}||\vec{\ell}||_1=\mathfrak{m}}}. \tag{9.17}$$

It is easy to see

$$\sharp\{\vec{\ell}\in[\![0,4]\!]^W : ||\vec{\ell}||_1 = \mathfrak{m}\} \leq \binom{4W}{\mathfrak{m}}. \tag{9.18}$$

Moreover, it is obvious that

$$\sharp\{\vec{\boldsymbol{\alpha}},\vec{\boldsymbol{\beta}}\in[\![1,4]\!]^{||\vec{\ell}||_1}\} = 16^{||\vec{\ell}||_1}. \tag{9.19}$$

Therefore, it suffices to investigate the integral

$$\mathfrak{P}_{\vec{\ell},\vec{\boldsymbol{\alpha}},\vec{\boldsymbol{\beta}}} := \int \mathrm{d}\Omega \mathrm{d}\Xi \exp\Big\{ -\sum_{j,k} \tilde{\mathfrak{s}}_{jk} Tr\Omega_j \Xi_k - \sum_{j=1}^{W} TrD_{\pm}^{-1}\Omega_j D_{\pm}^{-1}\Xi_j \Big\} \prod_{j=1}^{W}\prod_{i=1}^{\ell_j} \omega_{j,\alpha_{j,i}} \xi_{j,\beta_{j,i}}$$

for each combination $(\vec{\ell},\vec{\boldsymbol{\alpha}},\vec{\boldsymbol{\beta}})$, and then sum it up for $(\vec{\ell},\vec{\boldsymbol{\alpha}},\vec{\boldsymbol{\beta}})$ to get the estimate of $\mathsf{P}(\hat{X},\hat{B},V,T)$. Specifically, we have the following lemma.

Lemma 9.3. *With the notation above, we have*

$$\mathfrak{P}_{\vec{\ell},\vec{\boldsymbol{\alpha}},\vec{\boldsymbol{\beta}}} = 0, \quad \text{if} \quad ||\vec{\ell}||_1 = 0 \quad \text{or} \quad 1. \tag{9.20}$$

Moreover, we have

$$|\mathfrak{P}_{\vec{\ell},\vec{\boldsymbol{\alpha}},\vec{\boldsymbol{\beta}}}| \le |\det \mathbb{A}_+|^2 \det(S^{(1)})^2 \big(||\vec{\ell}||_1 - 1\big)! (2W^{\gamma})^{(||\vec{\ell}||_1 - 1)}, \quad \text{if} \quad ||\vec{\ell}||_1 \ge 2. \tag{9.21}$$

We postpone the proof of Lemma 9.3 and prove Lemma 9.2 at first.

Proof of Lemma 9.2. By (4.4), (9.14) and (9.20) and the fact that $\hat{\mathfrak{p}}_0(\cdot) = O(1)$, we have

$$|\mathsf{P}(\hat{X},\hat{B},V,T)| \le C \sum_{\substack{\vec{\ell}\in[\![0,4]\!]^W, \\ \text{s. t.} ||\vec{\ell}||_1 \ge 2}} M^{-\frac{||\vec{\ell}||_1}{2}} \sum_{\vec{\boldsymbol{\alpha}},\vec{\boldsymbol{\beta}} \in [\![1,4]\!]^{||\vec{\ell}||_1}} \prod_{j=1}^{W} |\mathring{\mathfrak{q}}_{\ell_j,j,\boldsymbol{\alpha}_j,\boldsymbol{\beta}_j}| \cdot |\mathfrak{P}_{\vec{\ell},\vec{\boldsymbol{\alpha}},\vec{\boldsymbol{\beta}}}|. \tag{9.22}$$

Substituting the bounds (9.16), (9.19) and (9.21) into (9.22) yields

$$\begin{aligned} |\mathsf{P}(\hat{X},\hat{B},V,T)| \le{}& |\det \mathbb{A}_+|^2 \det(S^{(1)})^2 \\ & \times \sum_{\substack{\vec{\ell}\in[\![0,4]\!]^W, \\ \text{s. t.} ||\vec{\ell}||_1 \ge 2}} e^{O(||\vec{\ell}||_1)} \cdot M^{-\frac{||\vec{\ell}||_1}{2}} \cdot \big(||\vec{\ell}||_1 - 1\big)! (2W^{\gamma})^{(||\vec{\ell}||_1 - 1)} \cdot \prod_{j=1}^{W} (1+\mathring{\kappa}_j)^{\ell_j}. \end{aligned} \tag{9.23}$$

Now, from (9.4) we have

$$\prod_{j=1}^{W} (1+\mathring{\kappa}_j)^{\ell_j} \le \Theta^{||\vec{\ell}||_1}, \tag{9.24}$$

which can absorb the irrelevant factor $e^{O(||\vec{\ell}||_1)}$. Using (9.17), (9.18) and (9.24), we have

$$\begin{aligned} & \sum_{\substack{\vec{\ell}\in[\![0,4]\!]^W, \\ \text{s.t.} ||\vec{\ell}||_1 \ge 2}} e^{O(||\vec{\ell}||_1)} \cdot M^{-\frac{||\vec{\ell}||_1}{2}} \cdot \big(||\vec{\ell}||_1 - 1\big)! (2W^{\gamma})^{(||\vec{\ell}||_1 - 1)} \cdot \prod_{j=1}^{W} (1+\mathring{\kappa}_j)^{\ell_j} \\ & \quad \le \sum_{\mathfrak{m}=2}^{4W} \binom{4W}{\mathfrak{m}} \cdot M^{-\frac{\mathfrak{m}}{2}} \cdot \Theta^{\mathfrak{m}} \cdot \mathfrak{m}! W^{(\mathfrak{m}-1)\gamma} \\ & \quad \le \sum_{\mathfrak{m}=2}^{4W} (4W)^{\mathfrak{m}} \cdot M^{-\frac{\mathfrak{m}}{2}} \cdot \Theta^{\mathfrak{m}} \cdot W^{(\mathfrak{m}-1)\gamma} = O\Big(\frac{W^{2+\gamma}\Theta^2}{M}\Big), \end{aligned} \tag{9.25}$$

where the last step follows from (5.33) and (5.38). Now, substituting (9.25) into (9.23), we can complete the proof of Lemma 9.2. □

Hence, what remains is to prove Lemma 9.3. We will need the following technical lemma whose proof is postponed.

Lemma 9.4. *For any index sets $\mathsf{I},\mathsf{J} \subset \{1,\dots,W\}$ with $|\mathsf{I}| = |\mathsf{J}| = \mathfrak{m} \ge 1$, we have the following bounds for the determinants of the submatrices of S, $\mathbb{A}_+$ and $\mathbb{A}_-$ defined in (8.14).*

- *For* $(\mathbb{A}_+)^{(\mathsf{I}|\mathsf{J})}$ *and* $(\mathbb{A}_-)^{(\mathsf{I}|\mathsf{J})}$*, we have*

$$\frac{|\det(\mathbb{A}_+)^{(\mathsf{I}|\mathsf{J})}|}{|\det \mathbb{A}_+|} \leq 1, \quad \frac{|\det(\mathbb{A}_-)^{(\mathsf{I}|\mathsf{J})}|}{|\det \mathbb{A}_-|} \leq 1. \tag{9.26}$$

- *For* $S^{(\mathsf{I}|\mathsf{J})}$*, we have*

$$\frac{|\det S^{(\mathsf{I}|\mathsf{J})}|}{|\det S^{(1)}|} \leq (\mathfrak{m}-1)!(2W^{\gamma})^{(\mathfrak{m}-1)}. \tag{9.27}$$

Proof of Lemma 9.3. Recall the definition in (6.20). Furthermore, we introduce the matrix

$$\mathbb{H} = (a_+^{-2}\mathbb{A}_+) \oplus S \oplus S \oplus (a_-^{-2}\mathbb{A}_-). \tag{9.28}$$

Using the fact $a_+a_- = -1$, we can write

$$-\sum_{j,k} \tilde{\mathfrak{s}}_{jk} Tr\Omega_j \Xi_k - \sum_{j=1}^{W} TrD_{\pm}^{-1}\Omega_j D_{\pm}^{-1}\Xi_j = -\vec{\Omega}\mathbb{H}\vec{\Xi}'.$$

Now, by the Gaussian integral of the Grassmann variables in (3.2), we see that

$$|\mathfrak{P}_{\vec{\ell},\vec{\boldsymbol{\alpha}},\vec{\boldsymbol{\beta}}}| = |\det \mathbb{H}^{(\mathsf{I}|\mathsf{J})}| \tag{9.29}$$

for some index sets $\mathsf{I}, \mathsf{J} \subset \{1, \ldots, 4W\}$ determined by $\vec{\boldsymbol{\alpha}}$ and $\vec{\boldsymbol{\beta}}$ such that

$$|\mathsf{I}| = |\mathsf{J}| = ||\vec{\ell}||_1.$$

Here we mention that (9.29) may fail when at least two components in $\boldsymbol{\alpha}_j$ coincide for some j. But $\mathfrak{P}_{\vec{\ell},\vec{\boldsymbol{\alpha}},\vec{\boldsymbol{\beta}}} = 0$ in this case because of $\chi^2 = 0$ for any Grassmann variable χ.

Now, obviously, there exists index sets $\mathsf{I}_\alpha, \mathsf{J}_\alpha \subset \{1, \ldots, W\}$ for $\alpha = 1, \ldots, 4$ such that

$$\mathbb{H}^{(\mathsf{I}|\mathsf{J})} = (a_+^{-2}\mathbb{A}_+)^{(\mathsf{I}_1|\mathsf{J}_1)} \oplus S^{(\mathsf{I}_2|\mathsf{J}_2)} \oplus S^{(\mathsf{I}_3|\mathsf{J}_3)} \oplus (a_-^{-2}\mathbb{A}_-)^{(\mathsf{I}_4|\mathsf{J}_4)}, \quad \sum_{\alpha} |\mathsf{I}_\alpha| = \sum_{\alpha} |\mathsf{J}_\alpha| = ||\vec{\ell}||_1.$$

It suffices to consider the case

$$|\mathsf{I}_\alpha| = |\mathsf{J}_\alpha|, \quad \forall\, \alpha = 1, 2, 3, 4.$$

Otherwise, $\det \mathbb{H}^{(\mathsf{I}|\mathsf{J})}$ is obviously 0, in light of the block structure of $\mathbb{H}$, see the definition (9.28). Now, note that, since $\det S = 0$, we have

$$\det \mathbb{H}^{(\mathsf{I}|\mathsf{J})} = 0, \qquad \text{if} \quad ||\vec{\ell}||_1 = 0, 1.$$

In addition, if $||\vec{\ell}||_1 = 2$, by using (9.30) below, one has

$$|\det \mathbb{H}^{(\mathsf{I}|\mathsf{J})}| = \begin{cases} |\det \mathbb{A}_+\mathbb{A}_-| \det(S^{(1)})^2, & \text{if} \quad |\mathsf{I}_2| = |\mathsf{I}_3| = |\mathsf{J}_2| = |\mathsf{J}_3| = 1, \\ 0, & \text{otherwise.} \end{cases}$$

For more general $\vec{\ell}$, by Lemma 9.4, we have

$$|\det \mathbb{H}^{(\mathsf{I}|\mathsf{J})}| \leq |\det \mathbb{A}_+\mathbb{A}_-| \det(S^{(1)})^2 \big(||\vec{\ell}||_1 - 1\big)!(2W^{\gamma})^{(||\vec{\ell}||_1 - 1)}.$$

Then, by the fact $|\det \mathbb{A}_+\mathbb{A}_-| = |\det \mathbb{A}_+|^2$, we can conclude the proof of Lemma 9.3. □

To prove Lemma 9.4, we will need the following lemma.

Lemma 9.5. *For the weighted Laplacian* S*, we have*

$$\det S^{(i|j)} = (-1)^{j-i} \det S^{(i)}, \quad \forall\, i, j = 1, \ldots, W \tag{9.30}$$

Remark 9.6. *A direct consequence of (9.30) is*

$$\det S^{(1)} = \ldots = \det S^{(W)}. \tag{9.31}$$

Proof of Lemma 9.5. Without loss of generality, we assume $j > i$ in the sequel. We introduce the matrices

$$P_{ij} := I_{i-1} \oplus \begin{pmatrix} & I_{j-i-1} \\ 1 & \end{pmatrix} \oplus I_{W-j}, \qquad E_j := I - 2\mathbf{e}_j\mathbf{e}_j^* - \sum_{\ell \neq j} \mathbf{e}_\ell \mathbf{e}_j^*.$$

It is not difficult to check

$$S^{(i|j)} = S^{(i)} P_{ij} E_j. \tag{9.32}$$

Then, by the fact $\det P_{ij}E_j = (-1)^{j-i}$, we can get the conclusion. □

Proof of Lemma 9.4. At first, by the definition in (8.14), (1.4) and the fact $\mathsf{Re}a_+^2 = \mathsf{Re}a_-^2 > 0$, it is easy to see that the singular values of $\mathbb{A}_+$ and $\mathbb{A}_-$ are all larger than 1. With the aid of the rectangular matrix $(\mathbb{A}_+)^{(\mathsf{I}|\emptyset)}$ as an intermediate matrix, we can use Cauchy interlacing property twice to see that the k-th largest singular value of $(\mathbb{A}_+)^{(\mathsf{I}|\mathsf{J})}$ is always smaller than the k-th largest singular value of $\mathbb{A}_+$. Consequently, we have the first inequality of (9.26). In the same manner, we can get the second inequality of (9.26)

Now, we prove (9.27). At first, we address the case that $\mathsf{I} \cap \mathsf{J} \neq \emptyset$. In light of (9.31), without loss of generality, we assume that $1 \in \mathsf{I} \cap \mathsf{J}$. Then $S^{(\mathsf{I}|\mathsf{J})}$ is a submatrix of $S^{(1)}$. Therefore, we can find two permutation matrices P and Q, such that

$$PS^{(1)}Q = \begin{pmatrix} \mathrm{A} & \mathrm{B} \\ \mathrm{C} & \mathrm{D} \end{pmatrix},$$

where $\mathrm{D} = S^{(\mathsf{I}|\mathsf{J})}$. Now, by Schur complement, we know that

$$\frac{|\det S^{(\mathsf{I}|\mathsf{J})}|}{|\det S^{(1)}|} = |\det(\mathrm{A} - \mathrm{B}\mathrm{D}^{-1}\mathrm{C})^{-1}|.$$

Moreover, $(\mathrm{A} - \mathrm{B}\mathrm{D}^{-1}\mathrm{C})^{-1}$ is the $(|\mathsf{I}|-1)$ by $(|\mathsf{I}|-1)$ upper-left corner of

$$(PS^{(1)}Q)^{-1} = Q^{-1}(S^{(1)})^{-1}P^{-1}.$$

That means $\det S^{(\mathsf{I}|\mathsf{J})}/\det S^{(1)}$ is the determinant of a sub matrix of $(S^{(1)})^{-1}$ (with dimension $|\mathsf{I}|-1$), up to a sign. Then, by Assumption 1.1 (iii), we can easily get

$$|\det S^{(\mathsf{I}|\mathsf{J})}|/|\det S^{(1)}| \leq (|\mathsf{I}|-1)!W^{(|\mathsf{I}|-1)\gamma}.$$

Now, for the case $\mathsf{I} \cap \mathsf{J} = \emptyset$, we can fix one $i \in \mathsf{I}$ and $j \in \mathsf{J}$. Due to (9.30), it suffices to consider

$$\frac{\det S^{(\mathsf{I}|\mathsf{J})}}{\det S^{(i|j)}}. \tag{9.33}$$

By similar discussion, one can see that (9.33) is the determinant of a sub matrix of $(S^{(i|j)})^{-1}$ with dimension $|\mathsf{I}|-1$. Hence, it suffices to investigate the bound of the entries of $(S^{(i|j)})^{-1}$. From (9.32) we have

$$(S^{(i|j)})^{-1} = E_j^{-1}P_{ij}^{-1}(S^{(i)})^{-1}. \tag{9.34}$$

Observe that

$$P_{ij}^{-1} = I_{i-1} \oplus \begin{pmatrix} & 1 \\ I_{j-i-1} & \end{pmatrix} \oplus I_{W-j}, \qquad E_j^{-1} = E_j.$$

Then, it is elementary to see that the entries of $(S^{(i|j)})^{-1}$ are bounded by $2W^\gamma$, in light of (9.34) and Assumption 1.1 (iii). Consequently, we have

$$\frac{|\det S^{(\mathsf{I}|\mathsf{J})}|}{|\det S^{(i|j)}|} \leq (|\mathsf{I}|-1)!(2W^\gamma)^{|\mathsf{I}|-1},$$

which implies (9.27). Hence, we completed the proof of Lemma 9.4. □

9.2. $\mathsf{F}(\hat{X},\hat{B},V,T)$ **in the Type I' vicinity.** Neglecting the $X^{[1]}$, $\mathbf{y}^{[1]}$ and $\mathbf{w}^{[1]}$-variables in $\mathsf{Q}(\cdot)$ at first, we investigate the integral $\mathsf{F}(\hat{X},\hat{B},V,T)$ in the Type I' vicinity in this section. We have the following lemma.

Lemma 9.7. *Suppose that the assumptions in Theorem 1.14 hold. In the Type I' vicinity, we have*

$$\mathsf{F}(\hat{X},\hat{B},V,T)=O\Big(\frac{1}{(N\eta)^{n+2}}\Big). \tag{9.35}$$

Recalling the functions $\mathbb{G}(\hat{B},T)$ and $\mathbb{F}(\hat{X},V)$ defined in (6.23) and (6.24), we further introduce

$$\mathring{\mathbb{G}}(\hat{B},T)=\exp\big\{(a_+-a_-)N\eta\big\}\mathbb{G}(\hat{B},T),\quad \mathring{\mathbb{F}}(\hat{X},V)=\exp\big\{-(a_+-a_-)N\eta\big\}\mathbb{F}(\hat{X},V). \tag{9.36}$$

Then, we have the decomposition

$$\mathsf{F}(\hat{X},\hat{B},V,T)=\mathring{\mathbb{G}}(\hat{B},T)\mathring{\mathbb{F}}(\hat{X},V). \tag{9.37}$$

Hence, we can estimate $\mathring{\mathbb{F}}(\hat{X},V)$ and $\mathring{\mathbb{G}}(\hat{B},T)$ separately in the sequel.

9.2.1. *Estimate of* $\mathring{\mathbb{F}}(\hat{X},V)$. We have the following lemma.

Lemma 9.8. *Suppose that the assumptions in Theorem 1.14 hold. In the Type I' vicinity, we have*

$$\mathring{\mathbb{F}}(\hat{X},V)=O\Big(\frac{1}{N\eta}\Big). \tag{9.38}$$

Proof. Using (8.8) and (8.9), we can write

$$X_j=P_1^*V_j^*\hat{X}_jV_jP_1=P_1^*D_\pm P_1+O\Big(\frac{\Theta}{\sqrt{M}}\Big), \tag{9.39}$$

where the remainder term represents a 2×2 matrix whose max-norm is bounded by $\Theta/\sqrt{M}$. Using (9.39) and recalling $N=MW$ yields

$$\exp\Big\{M\eta\sum_{j=1}^{W}TrX_jJ\Big\}=\exp\Big\{N\eta TrP_1^*D_\pm P_1J\Big\}\Big(1+O\Big(\frac{\Theta N\eta}{\sqrt{M}}\Big)\Big). \tag{9.40}$$

Substituting (9.40) into (3.19) and (6.24), we can write

$$\begin{aligned}\mathbb{F}(\hat{X},V)=&\int \mathrm{d}\mu(P_1)\mathrm{d}X^{[1]}\ \exp\Big\{N\eta TrP_1^*D_\pm P_1J\Big\}\\&\times\prod_{k=p,q}\frac{1}{\det{}^2(X_k^{[1]})}\cdot\exp\Big\{\mathbf{i}TrX_k^{[1]}JZ-\sum_j\tilde{\mathfrak{s}}_{jk}TrX_jX_k^{[1]}J\Big\}\\&\times\prod_{k,\ell=p,q}\exp\Big\{\frac{\tilde{\mathfrak{s}}_{k\ell}}{2M}TrX_k^{[1]}JX_\ell^{[1]}J\Big\}\cdot\Big(1+O\Big(\frac{\Theta N\eta}{\sqrt{M}}\Big)\Big).\end{aligned}$$

Recalling the parametrization of P_1 in (3.24), we have

$$TrP_1^*D_\pm P_1J=(1-2v^2)(a_+-a_-).$$

Consequently, we have

$$\begin{aligned}\mathring{\mathbb{F}}(\hat{X},V)=&\int\mathrm{d}X^{[1]}\int v\mathrm{d}v\int\frac{\mathrm{d}\theta}{\pi}\ \exp\Big\{-2(a_+-a_-)N\eta v^2\Big\}\prod_{k=p,q}\frac{1}{\det{}^2(X_k^{[1]})}\\&\times\prod_{k=p,q}\exp\Big\{\mathbf{i}TrX_k^{[1]}JZ-\sum_j\tilde{\mathfrak{s}}_{jk}TrX_jX_k^{[1]}J\Big\}(1+o(1)).\end{aligned}$$

Obviously, by the fact that $X^{[1]}$-variables are all bounded and $|\det X_k^{[1]}| = 1$ for $k = p, q$, it is easy to see that

$$|\mathring{\mathbb{F}}(\hat{X}, V)| \le C \int_0^1 v\mathrm{d}v \ \exp\Big\{ -2(a_+ - a_-)N\eta v^2 \Big\} = O\Big(\frac{1}{N\eta}\Big).$$

Therefore, we completed the proof. □

9.2.2. *Estimate of* $\mathring{\mathbb{G}}(\hat{B}, T)$. Recall the definition of $\mathring{\mathbb{G}}(\hat{B}, T)$ from (9.36), (6.23) and (3.20). In this section, we will prove the following lemma.

Lemma 9.9. *Suppose that the assumptions in Theorem 1.14 hold. In the Type I' vicinity, we have*

$$\mathring{\mathbb{G}}(\hat{B}, T) = O\Big(\frac{1}{(N\eta)^{n+1}}\Big). \tag{9.41}$$

Note that $y_p^{[1]}$, $y_q^{[1]}$ and t in the parametrization of Q_1 (see (3.24)) are not bounded, we shall truncate them with some appropriate bounds at first, whereby we can neglect some irrelevant terms in the integrand, in order to simplify the integral. More specifically, we will do the truncations

$$t \le \frac{1}{(N\eta)^{1/4}} \tag{9.42}$$

and

$$y_p^{[1]}, y_q^{[1]} \le (N\eta)^{\frac{1}{8}}. \tag{9.43}$$

Accordingly, we set

$$\begin{aligned}\widehat{\mathbb{G}}(\hat{B}, T) := e^{(a_+ - a_-)N\eta} \int_{\mathbb{L}} \frac{\mathrm{d}\sigma}{2\pi} \int_{\mathbb{I}^2} v_p^{[1]} v_q^{[1]} \mathrm{d}v_p^{[1]} \mathrm{d}v_q^{[1]} \int_0^{(N\eta)^{\frac{1}{8}}} \mathrm{d}y_p^{[1]} \int_0^{(N\eta)^{\frac{1}{8}}} \mathrm{d}y_q^{[1]} \\ \times \int_0^{(N\eta)^{-\frac{1}{4}}} 2t\mathrm{d}t \int_{\mathbb{L}^2} \mathrm{d}\sigma_p^{[1]} \mathrm{d}\sigma_q^{[1]} \ g(Q_1, T, \hat{B}, \mathbf{y}^{[1]}, \mathbf{w}^{[1]}),\end{aligned} \tag{9.44}$$

where we have used the parameterization of $\mathbf{w}^{[1]}$ in (3.15). We will prove the following lemma.

Lemma 9.10. *Suppose that the assumptions in Theorem 1.14 hold. In the Type I' vicinity, we have*

$$\mathring{\mathbb{G}}(\hat{B}, T) = \widehat{\mathbb{G}}(\hat{B}, T) + O(e^{-N^{\varepsilon}})$$

for some positive constant ε.

Proof. At first, by (6.29)-(6.32), we have for any j,

$$\mathsf{Re} Tr B_j Y_k^{[1]} J \ge \frac{y_k^{[1]}}{(s+t)^2} \cdot \frac{\min\{\mathsf{Re} b_{j,1}, \mathsf{Re} b_{j,2}\}}{(s_j + t_j)^2} \ge c \frac{y_k^{[1]}}{1 + 2t^2}, \quad k = p, q, \tag{9.45}$$

for some positive constant c, where the last step follows from the facts that $\mathsf{Re} b_{j,1}, \mathsf{Re} b_{j,2} = \mathsf{Re} a_+ + o(1)$ and $t_j = o(1)$ in the Type I' vicinity. In addition, it is not difficult to get

$$Tr B_j J = \Big(a_+ - a_- + O\Big(\frac{\Theta}{\sqrt{M}}\Big)\Big)(1 + 2t^2), \quad \forall\, j = 1, \dots, W,$$

which implies that

$$M\eta \sum_{j=1}^{W} Tr B_j J = (a_+ - a_-)N\eta + 2\Big(a_+ - a_- + O\Big(\frac{\Theta}{\sqrt{M}}\Big)\Big) N\eta t^2 + O\Big(\frac{\Theta N\eta}{\sqrt{M}}\Big). \tag{9.46}$$

Note that the second and third factors in the definition of $g(\cdot)$ in (3.20) can be bounded by 1, according to (6.26). Then, as a consequence of (9.45) and (9.46), we have

$$e^{(a_+ - a_-)N\eta} |g(\cdot)| \le C (y_p^{[1]} y_q^{[1]})^{n+3} \exp\{-c' N\eta t^2\} \exp\Big\{ -c \frac{y_p^{[1]} + y_q^{[1]}}{1 + 2t^2} \Big\}, \tag{9.47}$$

for some positive constants C, c and c'. By integrating $y_p^{[1]}$ and $y_q^{[1]}$ out at first, we can easily see that the truncation (9.42) only produces an error of order $O(\exp\{-N^{\varepsilon}\})$ to the integral $\mathring{\mathbb{G}}(\hat{B},T)$, for some positive constant $\varepsilon=\varepsilon(\varepsilon_2)$ by the assumption $\eta\geq N^{-1+\varepsilon_2}$ in (1.15). Then one can substitute the bound (9.42) to the last factor of the r.h.s. of (9.47), thus

$$\exp\Big\{-c\frac{y_p^{[1]}+y_q^{[1]}}{1+2t^2}\Big\}\leq\exp\Big\{-\frac{c}{2}(y_p^{[1]}+y_q^{[1]})\Big\}.$$

We can also do the truncation (9.43) in the integral $\mathring{\mathbb{G}}(\hat{B},T)$, up to an error of order $O(\exp\{-N^{\varepsilon}\})$, for some positive constant ε. Therefore, we completed the proof of Lemma 9.10. □

With the aid of Lemma 9.10, it suffices to work on $\widehat{\mathbb{G}}(\hat{B},T)$ in the sequel. We have the following lemma.

Lemma 9.11. *We have*

$$\widehat{\mathbb{G}}(\hat{B},T)=O\Big(\frac{1}{(N\eta)^{n+1}}\Big).$$

Proof of Lemma 9.11. Recall the parameterization of $\mathbf{w}_k^{[1]}$ in (3.15) again. To simplify the notation, we set

$$\mathfrak{w}_k^{[1]}=u_k^{[1]}v_k^{[1]},\quad k=p,q.$$

Similarly to (9.39), using $t=o(1)$ from (9.42), we have the expansion

$$B_j=Q_1^{-1}T_j^{-1}\hat{B}_jT_jQ_1=Q_1^{-1}D_{\pm}Q_1+O\Big(\frac{\Theta}{\sqrt{M}}\Big).$$

Consequently, we have

$$-M\eta\sum_{j=1}^{W}TrB_jJ=-N\eta(a_+-a_-)(1+2t^2)+O\Big(\frac{\Theta N\eta}{\sqrt{M}}\Big).\tag{9.48}$$

In addition, for $k=p,q$, using the fact $\sum_j\tilde{\mathfrak{s}}_{jk}=1$, we have

$$\begin{aligned}\sum_j\tilde{\mathfrak{s}}_{jk}TrB_jY_k^{[1]}J&=TrQ_1^{-1}D_{\pm}Q_1Y_k^{[1]}J+\frac{\Theta}{\sqrt{M}}TrR_kY_k^{[1]}\\&=y_k^{[1]}\Big((a_+-a_-)t^2+\big(a_+(u_k^{[1]})^2-a_-(v_k^{[1]})^2\big)\Big)\\&\quad+y_k^{[1]}\Big((a_+-a_-)\big(e^{-\mathbf{i}(\sigma_k^{[1]}+\sigma)}+e^{\mathbf{i}(\sigma_k^{[1]}+\sigma)}\big)\mathfrak{w}_k^{[1]}st\Big)+\frac{\Theta}{\sqrt{M}}TrR_kY_k^{[1]},\end{aligned}\tag{9.49}$$

where R_k is a 2×2 matrix independent of $Y_k^{[1]}$, satisfying

$$||R_k||_{\max}=O(1).\tag{9.50}$$

Observe that the term in (9.48) is obviously independent of $\mathbf{w}^{[1]}$-variables. In addition, for $k=p$ or q, we have

$$\mathbf{i}TrY_k^{[1]}JZ=\big(-\eta+\mathbf{i}E(1-2(v_k^{[1]})^2)\big)y_k^{[1]},\tag{9.51}$$

and for $k,\ell=p$ or q, we have

$$TrY_k^{[1]}JY_\ell^{[1]}J=y_k^{[1]}y_\ell^{[1]}\Big((\mathfrak{w}_k^{[1]}\mathfrak{w}_\ell^{[1]})^2+\mathfrak{w}_k^{[1]}\mathfrak{w}_\ell^{[1]}\Big(e^{\mathbf{i}(\sigma_k^{[1]}-\sigma_\ell^{[1]})}-e^{\mathbf{i}(\sigma_\ell^{[1]}-\sigma_k^{[1]})}\Big)\Big).\tag{9.52}$$

Moreover, we have

$$\Big(\Big(\mathbf{w}_q^{[1]}(\mathbf{w}_q^{[1]})^*\Big)_{12}\Big(\mathbf{w}_p^{[1]}(\mathbf{w}_p^{[1]})^*\Big)_{21}\Big)^n=\Big(\mathfrak{w}_p^{[1]}\mathfrak{w}_q^{[1]}\Big)^ne^{\mathbf{i}n(\sigma_p^{[1]}-\sigma_q^{[1]})}.\tag{9.53}$$

Substituting (9.48), (9.49) and (9.51)-(9.53) to the definition of $g(\cdot)$ in (3.20) and reordering the factors properly, we can write the integrand in (9.44) as

$$\begin{aligned}
&\exp\{(a_+-a_-)N\eta\}g(\cdot)\\
&=\exp\{\mathbf{i}n(\sigma_p^{[1]}-\sigma_q^{[1]})\}\cdot\exp\Big\{-(a_+-a_-)st\sum_{k=p,q}y_k^{[1]}\mathfrak{w}_k^{[1]}\Big(e^{-\mathbf{i}(\sigma_k^{[1]}+\sigma)}+e^{\mathbf{i}(\sigma_k^{[1]}+\sigma)}\Big)\Big\}\\
&\times\exp\Big\{-\frac{\Theta}{\sqrt{M}}\sum_{k=p,q}TrR_kY_k^{[1]}\Big\}\cdot\exp\Big\{-\frac{1}{M}\tilde{\mathfrak{s}}_{pq}y_p^{[1]}y_q^{[1]}\mathfrak{w}_p^{[1]}\mathfrak{w}_q^{[1]}\Big(e^{\mathbf{i}(\sigma_p^{[1]}-\sigma_q^{[1]})}-e^{\mathbf{i}(\sigma_q^{[1]}-\sigma_p^{[1]})}\Big)\Big\}\\
&\times\prod_{k=p,q}(y_k^{[1]})^{n+3}(\mathfrak{w}_k^{[1]})^n\cdot\prod_{k,\ell=p,q}\exp\Big\{-\frac{1}{2M}\tilde{\mathfrak{s}}_{k\ell}y_k^{[1]}y_\ell^{[1]}\Big(\mathfrak{w}_k^{[1]}\mathfrak{w}_\ell^{[1]}\Big)^2\Big\}\cdot\exp\big\{-2N\eta(a_+-a_-)t^2\big\}\\
&\times\prod_{k=p,q}\exp\Big\{-y_k^{[1]}\Big(\Big(a_+(u_k^{[1]})^2-a_-(v_k^{[1]})^2\Big)+(a_+-a_-)t^2+\eta-\mathbf{i}E\Big(1-2(v_k^{[1]})^2\Big)\Big)\Big\}\\
&\times\Big(1+O\Big(\frac{\Theta N\eta}{\sqrt{M}}\Big)\Big),
\end{aligned}\tag{9.54}$$

where the last factor is independent of the $\mathbf{w}^{[1]}$-variables. Here, we put the factors containing $\sigma_p^{[1]}$ and $\sigma_q^{[1]}$ together, namely, the first two lines on the r.h.s. of (9.54).

For further discussion, we write for $k=p,q$

$$TrR_kY_k^{[1]}=y_k^{[1]}\big(\mathfrak{r}_k^+e^{\mathbf{i}\sigma_k}+\mathfrak{r}_k^-e^{-\mathbf{i}\sigma_k}+\mathfrak{r}_k\big),\tag{9.55}$$

where $\mathfrak{r}_k^+$, $\mathfrak{r}_k^-$ and $\mathfrak{r}_k$ are all polynomials of $u_k^{[1]}$ and $v_k^{[1]}$, with bounded degree and bounded coefficients, in light of (9.50), the definition of $Y_k^{[1]}$ in (3.14) and the parametrization in (3.15).

Now, we start to estimate the integral (9.44) by using (9.54). We deal with the integral over $\sigma_p^{[1]}$ and $\sigma_q^{[1]}$ at first. These variables are collected in the integral of the form

$$\begin{aligned}
\mathcal{I}_\sigma(\ell_1,\ell_2):=\int_{\mathbb{L}^2}\mathrm{d}\sigma_p^{[1]}\mathrm{d}\sigma_q^{[1]}\exp\big\{\mathbf{i}(n+\ell_1)\sigma_p^{[1]}\big\}\exp\big\{-\mathbf{i}(n+\ell_2)\sigma_q^{[1]}\big\}\exp\Big\{-\frac{\Theta}{\sqrt{M}}\sum_{k=p,q}TrR_kY_k^{[1]}\Big\}\\
\times\exp\Big\{-(a_+-a_-)st\sum_{k=p,q}y_k^{[1]}\mathfrak{w}_k^{[1]}\big(e^{-\mathbf{i}(\sigma_k^{[1]}+\sigma)}+e^{\mathbf{i}(\sigma_k^{[1]}+\sigma)}\big)\Big\}\\
\times\exp\Big\{-\frac{1}{M}\tilde{\mathfrak{s}}_{pq}y_p^{[1]}y_q^{[1]}\mathfrak{w}_p^{[1]}\mathfrak{w}_q^{[1]}\big(e^{\mathbf{i}(\sigma_p^{[1]}-\sigma_q^{[1]})}-e^{\mathbf{i}(\sigma_q^{[1]}-\sigma_p^{[1]})}\big)\Big\}
\end{aligned}$$

with integers ℓ_1 and ℓ_2 independent of n. Note that according to (9.54), it suffices to consider $\mathcal{I}_\sigma(0,0)$ for the proof of (9.41). We study $\mathcal{I}_\sigma(\ell_1,\ell_2)$ for general ℓ_1 and ℓ_2 here, which will be used later.

Now, we set

$$\begin{aligned}
c_{p,q}&:=\tilde{\mathfrak{s}}_{pq}y_p^{[1]}y_q^{[1]}\mathfrak{w}_p^{[1]}\mathfrak{w}_q^{[1]},\\
c_{k,1}&:=-(a_+-a_-)sty_k^{[1]}\mathfrak{w}_k^{[1]}e^{-\mathbf{i}\sigma}-\frac{\Theta}{\sqrt{M}}y_k^{[1]}\mathfrak{r}_k^-,\quad k=p,q,\\
c_{k,2}&:=-(a_+-a_-)sty_k^{[1]}\mathfrak{w}_k^{[1]}e^{\mathbf{i}\sigma}-\frac{\Theta}{\sqrt{M}}y_k^{[1]}\mathfrak{r}_k^+,\quad k=p,q.
\end{aligned}\tag{9.56}$$

In addition, we introduce

$$d_{p,q}:=y_p^{[1]}y_q^{[1]},\quad d_k:=\Big(t+\frac{\Theta}{\sqrt{M}}\Big)y_k^{[1]},\quad k=p,q.\tag{9.57}$$

Obviously, when (9.42) is satisfied, we have

$$c_{p,q}=O(d_{p,q}),\quad c_{k,1}=O(d_k),\quad c_{k,2}=O(d_k),\quad k=p,q.\tag{9.58}$$

With the aid of the notation defined in (9.55) and (9.56), we can write

$$\mathcal{I}_\sigma(\ell_1,\ell_2)=\exp\Big\{-\frac{\Theta}{\sqrt{M}}(y_p^{[1]}\mathfrak{r}_p+y_q^{[1]}\mathfrak{r}_q)\Big\}\int_{\mathbb{L}^2}\mathrm{d}\sigma_p\mathrm{d}\sigma_q\ \exp\Big\{\mathbf{i}(n+\ell_1)\sigma_p^{[1]}\Big\}\exp\Big\{-\mathbf{i}(n+\ell_2)\sigma_q^{[1]}\Big\}$$
$$\times\prod_{k=p,q}\exp\Big\{c_{k,1}e^{-\mathbf{i}\sigma_k^{[1]}}+c_{k,2}e^{\mathbf{i}\sigma_k^{[1]}}\Big\}\exp\Big\{-\frac{c_{p,q}}{M}e^{\mathbf{i}(\sigma_p^{[1]}-\sigma_q^{[1]})}+\frac{c_{p,q}}{M}e^{\mathbf{i}(\sigma_q^{[1]}-\sigma_p^{[1]})}\Big\}. \tag{9.59}$$

We have the following lemma.

Lemma 9.12. *Under the truncation (9.42) and (9.43), we have*

$$|\mathcal{I}_\sigma(\ell_1,\ell_2)|\le C\Big(\Big(\frac{d_{p,q}}{M}\Big)^{n+\ell_3}+d_p^{2(n+\ell_3)}+d_q^{2(n+\ell_3)}\Big),\qquad \ell_3:=\frac{\ell_1+\ell_2}{2}$$

for some positive constant C.

Proof. At first, by Taylor expansion, we have

$$\exp\Big\{\mathbf{i}(n+\ell_1)\sigma_p^{[1]}\Big\}\exp\Big\{-\mathbf{i}(n+\ell_2)\sigma_q^{[1]}\Big\}\exp\Big\{-\frac{c_{p,q}}{M}e^{\mathbf{i}(\sigma_p^{[1]}-\sigma_q^{[1]})}+\frac{c_{p,q}}{M}e^{\mathbf{i}(\sigma_q^{[1]}-\sigma_p^{[1]})}\Big\}$$
$$=\sum_{n_1,n_2=0}^{\infty}\frac{(-1)^{n_1}}{(n_1)!(n_2)!}\Big(\frac{c_{p,q}}{M}\Big)^{n_1+n_2}\exp\Big\{\mathbf{i}(n+\ell_1+n_1-n_2)\sigma_p^{[1]}\Big\}\exp\Big\{-\mathbf{i}(n+\ell_2+n_1-n_2)\sigma_q^{[1]}\Big\}. \tag{9.60}$$

Now, for any $m_1,m_2\in\mathbb{Z}$, we denote

$$\tilde{\mathcal{I}}_\sigma(m_1,m_2):=\int_{\mathbb{L}^2}\mathrm{d}\sigma_p^{[1]}\mathrm{d}\sigma_q^{[1]}\ \exp\{\mathbf{i}m_1\sigma_p^{[1]}\}\exp\{-\mathbf{i}m_2\sigma_q^{[1]}\}\prod_{k=p,q}\exp\Big\{c_{k,1}e^{-\mathbf{i}\sigma_k^{[1]}}+c_{k,2}e^{\mathbf{i}\sigma_k^{[1]}}\Big\}$$
$$=4\pi^2\sum_{n_3=0}^{\infty}\mathbf{1}(n_3+m_1\ge0)\frac{(c_{p,1})^{n_3+m_1}(c_{p,2})^{n_3}}{n_3!(n_3+m_1)!}\sum_{n_4=0}^{\infty}\mathbf{1}(n_4+m_2\ge0)\frac{(c_{q,1})^{n_4}(c_{q,2})^{n_4+m_2}}{n_4!(n_4+m_2)!}. \tag{9.61}$$

Setting

$$m_1:=n+\ell_1+n_1-n_2,\qquad m_2:=n+\ell_2+n_1-n_2, \tag{9.62}$$

and using (9.60), we can rewrite (9.59) as

$$\mathcal{I}_\sigma(\ell_1,\ell_2)=\exp\Big\{-\frac{\Theta}{\sqrt{M}}(y_p^{[1]}\mathfrak{r}_p+y_q^{[1]}\mathfrak{r}_q)\Big\}\sum_{n_1,n_2=0}^{\infty}\frac{(-1)^{n_1}}{(n_1)!(n_2)!}\Big(\frac{c_{p,q}}{M}\Big)^{n_1+n_2}\tilde{\mathcal{I}}_\sigma(m_1,m_2). \tag{9.63}$$

For simplicity, we employ the notation

$$m_3:=m_3(\ell_1,n_1,n_2,n_3)=m_1+n_3,\quad m_4:=m_4(\ell_2,n_1,n_2,n_4)=m_2+n_4. \tag{9.64}$$

Consequently, by (9.63) and (9.61) we obtain

$$|\mathcal{I}_\sigma(\ell_1,\ell_2)|\le4\pi^2\Big|\exp\Big\{-\frac{\Theta}{\sqrt{M}}(y_p^{[1]}\mathfrak{r}_p+y_q^{[1]}\mathfrak{r}_q)\Big\}\Big|\sum_{n_1,n_2=0}^{\infty}\frac{1}{(n_1)!(n_2)!}\Big|\frac{c_{p,q}}{M}\Big|^{n_1+n_2}$$
$$\times\sum_{n_3=0}^{\infty}\mathbf{1}(m_3\ge0)\frac{|c_{p,1}|^{m_3}|c_{p,2}|^{n_3}}{n_3!m_3!}\cdot\sum_{n_4=0}^{\infty}\mathbf{1}(m_4\ge0)\frac{|c_{q,1}|^{n_4}|c_{q,2}|^{m_4}}{n_4!m_4!}$$
$$\le C\max_{n_1,n_2,n_3,n_4,m_3,m_4\in\mathbb{N}}\Big|\Big(\frac{c_{p,q}}{M}\Big)^{n_1+n_2}(c_{p,1})^{m_3}(c_{p,2})^{n_3}(c_{q,1})^{n_4}(c_{q,2})^{m_4}\Big|$$
$$\le C\max_{n_1,n_2,n_3,n_4,m_3,m_4\in\mathbb{N}}\Big|\Big(\frac{c_{p,q}}{M}\Big)^{n_1+n_2}(c_{p,1})^{m_3}(c_{q,2})^{m_4}\Big| \tag{9.65}$$

for some positive constant C, where in the last step we used the fact

$$|c_{k,1}|<1,\quad |c_{k,2}|<1,\quad k=p,q \tag{9.66}$$

which can be seen directly from the definition in (9.58), the truncations (9.42) and (9.43) and the assumption $\eta \le M^{-1}N^{\varepsilon_2}$. Analogously, we also have

$$\Big|\frac{c_{p,q}}{M}\Big| < 1. \tag{9.67}$$

According to the definitions (9.62) and (9.64), we have

$$2(n_1+n_2)+m_3+m_4 \ge 2n+\ell_1+\ell_2.$$

Hence, by using (9.66) and (9.67), we have the trivial bound

$$\max_{n_1,n_2,n_3,n_4,m_3,m_4\ge 0}\Big|\Big(\sqrt{\frac{|c_{p,q}|}{M}}\Big)^{2(n_1+n_2)}(c_{p,1})^{m_3}(c_{q,2})^{m_4}\Big| \le \Big(\frac{|c_{p,q}|}{M}\Big)^{n+\ell_3}+|c_{p,1}|^{2(n+\ell_3)}+|c_{q,2}|^{2(n+\ell_3)}.$$

Therefore, we completed the proof by using (9.58). □

Now, we return to the proof of Lemma 9.11. Using (9.54) and Lemma 9.12 with $\ell_1=\ell_2=0$ to (9.44), and integrating the bounded variables $v_p^{[1]}$, $v_q^{[1]}$ and σ out, we can get

$$\begin{aligned}|\widehat{\mathbb{G}}(\hat B,T)| \le C\int_0^{(N\eta)^{\frac18}}\mathrm{d}y_p^{[1]}\int_0^{(N\eta)^{\frac18}}\mathrm{d}y_q^{[1]}\int_0^{(N\eta)^{-\frac14}}2t\mathrm{d}t\cdot(d_{p,q})^{n+3}\\ \times\Big(\Big(\frac{d_{p,q}}{M}\Big)^n+d_p^{2n}+d_q^{2n}\Big)\exp\big\{-2N\eta(a_+-a_-)t^2\big\}\\ \times\exp\Big\{-\frac{\sqrt{4-E^2}}{2}\sum_{k=p,q}y_k^{[1]}\Big\}(1+o(1))\end{aligned}$$

where the last two factors come from the facts

$$\begin{aligned}&\Big|\exp\Big\{-\sum_{k=p,q}y_k^{[1]}\Big(a_+(u_k^{[1]})^2-a_-(v_k^{[1]})^2\Big)\Big\}\Big| = \exp\Big\{-\frac{\sqrt{4-E^2}}{2}\sum_{k=p,q}y_k^{[1]}\Big\},\\ &\Big|\exp\Big\{-\sum_{k=p,q}y_k^{[1]}\Big((a_+-a_-)t^2+\eta-\mathbf{i}E(1-2(v_k^{[1]})^2)\Big)\Big\}\Big(1+O\Big(\frac{\Theta N\eta}{\sqrt{M}}\Big)\Big)\Big| = 1+o(1).\end{aligned} \tag{9.68}$$

In (9.68) we used the fact $(u_k^{[1]})^2+(v_k^{[1]})^2=1$. Now, we integrate $y_p^{[1]}$ and $y_q^{[1]}$ out. Consequently, by the definition in (9.57), we have

$$|\widehat{\mathbb{G}}(\hat B,T)| \le C\int_0^{(N\eta)^{-\frac14}}2t\mathrm{d}t\,\Big(\frac{1}{M^n}+\Big(\frac{\Theta}{\sqrt{M}}\Big)^{2n}+t^{2n}\Big)\exp\big\{-2N\eta(a_+-a_-)t^2\big\} = O\Big(\frac{1}{(N\eta)^{n+1}}\Big),$$

where in the last step we have used the assumption $\eta \le M^{-1}N^{\varepsilon_2}$ in (1.15), Assumption 1.13, the definition of Θ in (5.33) and the fact $N=MW$. Hence, we completed the proof of Lemma 9.11. □

Finally, we can prove Lemma 9.9, and further prove Lemma 9.7.

Proof of Lemma 9.9 . This is a direct consequence of Lemmas 9.10 and 9.11. □

Proof of Lemma 9.7. This is a direct consequence of (9.37), Lemma 9.8 and Lemma 9.9. □

9.3. **Summing up: Proof of Lemma 9.1.** In this section, we slightly modify the discussions in Sections 9.1 and 9.2 to prove Lemma 9.1. The combination of Lemmas 9.2 and 9.7 would directly imply Lemma 9.1 if the $\mathsf{Q}(\cdot)$ factor were not present in the definition of $\mathsf{A}(\cdot)$. Now we should take $\mathsf{Q}(\cdot)$ into account. This argument is similar to the corresponding discussion in Section 6.4.

Proof of Lemma 9.1. At first, we observe that κ_1, κ_2 and κ_3 in (6.5) are obviously independent of n. Then, by the fact $\kappa_1 = W^{O(1)}$, it suffices to consider one monomial of the form

$$\mathfrak{p}_1\Big(t, s, (y_p^{[1]})^{-1}, (y_q^{[1]})^{-1}\Big)\mathfrak{p}_2\Big(\Big\{e^{\mathbf{i}\sigma_k^{[1]}}, e^{-\mathbf{i}\sigma_k^{[1]}}\Big\}_{k=p,q}\Big)\mathfrak{q}\Big(\Big\{\frac{\omega_{i,a}\xi_{j,b}}{M}\Big\}_{\substack{i,j=1,\ldots,W\\ a,b=1,\ldots,4}}\Big),$$

where the degrees of $\mathfrak{p}_1(\cdot)$, $\mathfrak{p}_2(\cdot)$ and $\mathfrak{q}(\cdot)$ are all $O(1)$, and independent of n, in light of the fact $\kappa_3 = O(1)$ in (6.5). Especially, the order of $(y_p^{[1]})^{-1}$and $(y_q^{[1]})^{-1}$ are not larger than 2, which can be easily seen from the definition of $\mathcal{Q}(\cdot)$ in (3.18).

Now, we reuse the notation $\mathsf{P}_\mathfrak{q}(\hat{X}, \hat{B}, V, T)$ and $\mathsf{F}_\mathfrak{p}(\hat{X}, \hat{B}, V, T)$ in (6.40), by redefining them as

$$\begin{aligned}\mathsf{P}_\mathfrak{q}(\hat{X}, \hat{B}, V, T) &:= \int \mathrm{d}\Omega\mathrm{d}\Xi\ \mathcal{P}(\cdot)\cdot\mathfrak{q}\Big(\Big\{\frac{\omega_{i,a}\xi_{j,b}}{M}\Big\}_{\substack{i,j=1,\ldots,W\\ a,b=1,\ldots,4}}\Big),\\ \mathsf{F}_\mathfrak{p}(\hat{X}, \hat{B}, V, T) &:= \int \mathrm{d}X^{[1]}\mathrm{d}\mathbf{y}^{[1]}\mathrm{d}\mathbf{w}^{[1]}\mathrm{d}\mu(P_1)\mathrm{d}\nu(Q_1)\ \mathcal{F}(\cdot)\\ &\quad\times\mathfrak{p}_1\Big(t, s, (y_p^{[1]})^{-1}, (y_q^{[1]})^{-1}\Big)\mathfrak{p}_2\Big(\Big\{e^{\mathbf{i}\sigma_k^{[1]}}, e^{-\mathbf{i}\sigma_k^{[1]}}\Big\}_{k=p,q}\Big).\end{aligned}$$

It is easy to check $\mathcal{P}(\cdot)\mathfrak{q}(\cdot)$ also has an expansion of the form in (9.14). Hence, $\mathsf{P}_\mathfrak{q}(\cdot)$ can be bounded in the same way as $\mathsf{P}(\cdot)$, thus we have

$$|\mathsf{P}_\mathfrak{q}(\hat{X}, \hat{B}, V, T)| \le \frac{W^{2+\gamma}\Theta^2}{M}|\det\mathbb{A}_+|^2\det(S^{(1)})^2.$$

For $\mathsf{F}_\mathfrak{p}(\cdot)$, the main modification is to use Lemma 9.12 with general ℓ_1 and ℓ_2 independent of n, owing to the function $\mathfrak{p}_2(\cdot)$. In addition, by the truncations in (9.42) and (9.43), we can bound $\mathfrak{p}_1(\cdot)$ by some constant C. Hence, it suffices to replace n by $n+\ell_3$ in the proof of Lemma 9.11. Finally, we can get

$$\mathsf{F}_\mathfrak{p}(\hat{X}, \hat{B}, V, T) = O\left(\frac{1}{(N\eta)^{n+\ell_3}}\right),$$

with some finite integer ℓ_3 independent of n. Consequently, we completed the proof of Lemma 9.1. □

10. Integral over the Type II and III vicinities

In this section, we prove Lemma 5.9. We only present the discussion for $\mathcal{I}(\Upsilon_+^b, \Upsilon_-^b, \Upsilon_+^x, \Upsilon_+^x, \Upsilon_S, \mathbb{I}^{W-1})$, i.e. integral over the Type II vicinity . The discussion on $\mathcal{I}(\Upsilon_+^b, \Upsilon_-^b, \Upsilon_-^x, \Upsilon_-^x, \Upsilon_S, \mathbb{I}^{W-1})$ is analogous. We start from (8.48). In this section, we will regard V-variables as fixed parameters, and consider the integrand as a function of all the other variables.

Similarly, we shall provide an estimate for the integrand. At first, under the parameterization (8.2) with $\varkappa = +$, we see that

$$x_{j,1} - x_{j,2} = \frac{\mathbf{i}a_+}{\sqrt{M}}(\mathring{x}_{j,1} - \mathring{x}_{j,2}) + O\Big(\frac{\Theta}{M}\Big), \qquad b_{j,1} + b_{j,2} = a_+ - a_- + O\Big(\frac{\Theta^{\frac{1}{2}}}{\sqrt{M}}\Big).$$

Consequently, we have

$$\begin{aligned}\prod_{j=1}^W (x_{j,1} - x_{j,2})^2(b_{j,1} + b_{j,2})^2 &= \frac{(-a_+^2)^W}{M^W}(a_+ - a_-)^{2W}\Big(1 + O\Big(\frac{\Theta^{\frac{3}{2}}}{\sqrt{M}}\Big)\Big)\\ &\quad\times\prod_{j=1}^W\Big(\mathring{x}_{j,1} - \mathring{x}_{j,2} + O\Big(\frac{\Theta}{\sqrt{M}}\Big)\Big)^2. \qquad (10.1)\end{aligned}$$

Now, analogously to the case of Type I vicinity, what remains is to estimate $\mathsf{A}(\hat{X}, \hat{B}, V, T)$. Our aim is to prove the following lemma.

Lemma 10.1. *Suppose that the assumptions in Theorem 1.14 hold. In the Type II vicinity, we have*

$$|\mathsf{A}(\hat{X},\hat{B},V,T)| \le e^{-cN\eta}|\det \mathbb{A}_+|^2 \det(S^{(1)})^2 \tag{10.2}$$

for some positive constant c.

With the aid of (10.1) and Lemma 10.1, we can prove Lemma 5.9.

Proof of Lemma 5.9. Recall (8.48). At first, by the definition of $\mathbb{A}_+^v$ in (8.35), (5.23) and the fact $\mathsf{Re}a_+^2 > 0$, we can see that

$$\mathsf{Re}(\mathring{\mathbf{x}}'\mathbb{A}_+^v\mathring{\mathbf{x}}) \ge ||\mathring{\mathbf{x}}||_2^2 \tag{10.3}$$

for all $\{V_j\}_{j=2}^W \in (\mathring{U}(2))^{W-1}$. Substituting (5.20), (10.1), (10.2), (10.3) and the estimates in Proposition 8.6 into (8.48) yields

$$\begin{aligned}
&|\mathcal{I}(\Upsilon_+^b,\Upsilon_-^b,\Upsilon_+^x,\Upsilon_+^x,\Upsilon_S,\mathbb{I}^{W-1})| \\
&\le e^{-cN\eta}\cdot\frac{(a_+-a_-)^{2W}}{8^W\pi^{3W-1}}\cdot|\det S^{(1)}|^2\cdot|\det\mathbb{A}_+|\cdot\int_{\mathbb{L}^{W-1}}\prod_{j=2}^W\frac{\mathrm{d}\theta_j}{2\pi}\int_{\mathbb{I}^{W-1}}\prod_{j=2}^W 2v_j\mathrm{d}v_j \\
&\times\int_{\mathbb{R}^{W-1}}\prod_{j=2}^W\mathrm{d}\tau_{j,1}\int_{\mathbb{R}^{W-1}}\prod_{j=2}^W\mathrm{d}\tau_{j,2}\int_{\mathbb{R}^W}\prod_{j=1}^W\mathrm{d}c_{j,1}\int_{\mathbb{R}^W}\prod_{j=1}^W\mathrm{d}c_{j,2}\int_{\mathbb{R}^W}\prod_{j=1}^W\mathrm{d}\mathring{x}_{j,1}\int_{\mathbb{R}^W}\prod_{j=1}^W\mathrm{d}\mathring{x}_{j,2} \\
&\times\exp\{(a_+-a_-)^2\boldsymbol{\tau}_1'S^{(1)}\boldsymbol{\tau}_1\}\exp\{(a_+-a_-)^2\boldsymbol{\tau}_2'S^{(1)}\boldsymbol{\tau}_2\}\exp\{-\frac{1}{2}||\mathring{\mathbf{c}}_1||_2^2-\frac{1}{2}||\mathring{\mathbf{c}}_2||_2^2\} \\
&\times\exp\{-\frac{1}{2}||\mathring{\mathbf{x}}_1||_2^2-\frac{1}{2}||\mathring{\mathbf{x}}_2||_2^2\}\prod_{j=1}^W\Big(\mathring{x}_{j,1}-\mathring{x}_{j,2}+O\Big(\frac{\Theta}{\sqrt{M}}\Big)\Big)^2,
\end{aligned}$$

where we absorbed several factors by $\exp\{-cN\eta\}$. We also enlarged the domains to the full ones. Then, using the trivial facts

$$\int_{\mathbb{L}^{W-1}}\prod_{j=2}^W\frac{\mathrm{d}\theta_j}{2\pi}\int_{\mathbb{I}^{W-1}}\prod_{j=2}^W 2v_j\mathrm{d}v_j = 1$$

and performing the Gaussian integral for the remaining variables, we can get

$$|\mathcal{I}(\Upsilon_+^b,\Upsilon_-^b,\Upsilon_+^x,\Upsilon_+^x,\Upsilon_S,\mathbb{I}^{W-1})| \le C|\det S^{(1)}|\cdot|\det\mathbb{A}_+|\cdot\Big(1+O\Big(\frac{\Theta}{\sqrt{M}}\Big)\Big)^W. \tag{10.4}$$

Observe that

$$|\det\mathbb{A}_+| \le |1+a_+^2|^W \le 2^W. \tag{10.5}$$

Moreover, by Assumption 1.1 (ii), we see that $|\mathfrak{s}_{ii}| \le (1-c_0)/2$ for some small positive constant c_0. Consequently, since $S^{(1)}$ is negative definite, we have

$$|\det S^{(1)}| \le \prod_{i\ne 1}|\mathfrak{s}_{ii}| \le \Big(\frac{1-c_0}{2}\Big)^W \tag{10.6}$$

by Hadamard's inequality. Substituting (10.5) and (10.6) into (10.4) yields

$$|\mathcal{I}(\Upsilon_+^b,\Upsilon_-^b,\Upsilon_+^x,\Upsilon_+^x,\Upsilon_S,\mathbb{I}^{W-1})| = O(e^{-cW}) \tag{10.7}$$

for some positive constant δ. Hence, we proved the first part of Lemma 5.9. The second part can be proved analogously. □

In the sequel, we prove Lemma 10.1. We also ignore the factor $\mathsf{Q}(\cdot)$ from the discussion at first.

10.1. $\mathsf{P}(\hat{X},\hat{B},V,T)$ **in the Type II vicinity.** Our aim, is to prove the following lemma.

Lemma 10.2. *Suppose that the assumptions in Theorem 1.14 hold. In the Type II vicinity, we have*

$$\mathsf{P}(\hat{X},\hat{B},V,T)\leq\frac{W^{2+\gamma}\Theta^2}{M}|\det\mathbb{A}_+|^2\det(S^{(1)})^2. \tag{10.8}$$

Proof. We will follow the strategy in Section 9.1. We regard all V-variables as fixed parameters. Now, we define the function

$$\mathring{\iota}\equiv\mathring{\iota}_j(\hat{X},\hat{B},T):=|\mathring{x}_{j,1}|+|\mathring{x}_{j,2}|+|\mathring{b}_{j,1}|+|\mathring{b}_{j,2}|+|\mathring{t}_j|.$$

Then, we recall the representation (9.5) and the definition of $\Delta_{\ell,j}$ in (9.6). We still adopt the representation (9.9). It is easy to see that in the Type II vicinity, we also have the bound (9.10) for $\mathring{\mathfrak{p}}_{\ell,j,\boldsymbol{\alpha},\boldsymbol{\beta}}$. The main difference is the first factor of the r.h.s. of (9.5). We expand it around the saddle point as

$$\exp\Big\{-TrV_j^*\hat{X}_j^{-1}V_j\Omega_jT_j^{-1}\hat{B}_j^{-1}T_j\Xi_j\Big\}=:\exp\Big\{-TrD_+^{-1}\Omega_jD_\pm^{-1}\Xi_j\Big\}\ \exp\Big\{-\frac{1}{\sqrt{M}}\widehat{\Delta}_j\Big\}.$$

We take the formula above as the definition of $\widehat{\Delta}_j$, which is of the form

$$\widehat{\Delta}_j=\sum_{\alpha,\beta=1}^4\hat{p}_{j,\alpha,\beta}\cdot\omega_{j,\alpha}\xi_{j,\beta},$$

where $\hat{p}_{j,\alpha,\beta}$ is a function of $\hat{X}$, $\hat{B}$, V and T-variables, satisfying

$$\hat{p}_{j,\alpha,\beta}=O(\mathring{\iota}).$$

Let

$$\widehat{\mathbb{H}}=(a_+^{-2}\mathbb{A}_+)\oplus S\oplus(a_+^{-2}\mathbb{A}_+)\oplus S.$$

Recalling the notation in (6.20), we can write

$$-\sum_{j,k}\tilde{\mathfrak{s}}_{jk}Tr\Omega_j\Xi_k-\sum_{j=1}^WTrD_+^{-1}\Omega_jD_\pm^{-1}\Xi_j=-\vec{\Omega}\widehat{\mathbb{H}}\vec{\Xi}'.$$

Now, via replacing $\Delta_{1,j}$ by $\widehat{\Delta}_{1,j}$, $\mathring{\kappa}_j$ by $\mathring{\iota}_j$, $\mathbb{H}$ by $\widehat{\mathbb{H}}$ in the proof of Lemma 9.2, we can perform the proof of Lemma 10.2 in the same way. We leave the details to the reader. □

10.2. $\mathsf{F}(\hat{X},\hat{B},V,T)$ **in the Type II vicinity.** In this section, we will prove the following lemma.

Lemma 10.3. *Suppose that the assumptions in Theorem 1.14 hold. In the Type II vicinity, we have*

$$\mathsf{F}(\hat{X},\hat{B},V,T)=O\Big(\frac{\exp\{-(a_+-a_-)N\eta\}}{(N\eta)^{n+1}}\Big). \tag{10.9}$$

Proof. Recall the decomposition (9.37). Note that Lemma 9.9 is still applicable. Hence, it suffices to estimate $\mathring{\mathbb{F}}(\hat{X},V)$. Now, note that in the Type II vicinity, it s obvious to see that

$$TrX_jJ=O\Big(\frac{|\mathring{x}_{j,1}|+|\mathring{x}_{j,2}|}{\sqrt{M}}\Big),\quad\forall\,j=1,\dots,W,$$

which implies that

$$\sum_{j=1}^WTrX_jJ=O\Big(\frac{||\mathring{\mathbf{x}}_1||_1+||\mathring{\mathbf{x}}_2||_1}{\sqrt{M}}\Big)=O\Big(\frac{\Theta}{\sqrt{M}}\Big).$$

Consequently, we have

$$\exp\Big\{M\eta\sum_{j=1}^WTrX_jJ\Big\}=\exp\{O(\Theta\sqrt{M}\eta)\}=1+o(1)$$

by our assumption on η. From (3.19) we can also see that all the other factors of $f(P_1, V, \hat{X}, X^{[1]})$ are $O(1)$. Hence, by the definition (9.36), we have

$$\mathring{\mathbb{F}}(\hat{X}, V) = O(\exp\{-(a_+ - a_-)N\eta\}),$$

which together with Lemma 9.9 yields the conclusion. □

10.3. **Summing up: Proof of Lemma 10.1.** Analogously, we shall slightly modify the proofs of Lemma 10.2 and Lemma 10.3, in order to take $\mathsf{Q}(\cdot)$ into account. The proof can then be performed in the same manner as Lemma 9.1. We omit the details here.

11. Proof of Theorem 1.14

The conclusion for Case 1 is a direct consequence of the discussions in Sections 3.5–10. The proofs of Case 2 and Case 3 can be performed analogously, with slight modifications, which will be stated below.

• (Case 1)

In this case, by using Lemmas 5.1, 5.6, 5.8 and 5.9, we can get (1.19) immediately.

• (Case 2)

In this case, we shall slightly modify the discussions in Sections 3.5-10 for Case 1, according to the decomposition of supermatrices in (3.9). Now, at first, in (3.12) and (3.13), for $A = \breve{\mathcal{S}}$, $\breve{X}$, $\breve{Y}$, $\breve{\Omega}$ or $\breve{\Xi}$, we replace $A_p^{\langle 1\rangle}$ and $A_q^{\langle 1\rangle}$ by $A_p^{\langle 1,2\rangle}$ and A_q respectively, and replace $A_q^{[1]}$ by $A_p^{[2]}$. In addition, in the last three lines of (3.13), we shall also replace $\tilde{s}_{jq}$ by $\tilde{s}_{jp}$, and replace $\tilde{s}_{pq}$ and $\tilde{s}_{qp}$ by $\tilde{s}_{pp}$, and in the first line, we replace $\bar{\phi}_{1,q,1}\phi_{1,p,1}\bar{\phi}_{2,p,1}\phi_{2,q,1}$ by $\bar{\phi}_{1,p,2}\phi_{1,p,1}\bar{\phi}_{2,p,1}\phi_{2,p,2}$. Then, in (3.14) and (3.15), for $A = X$, Y, Ω, Ξ, $\boldsymbol{\omega}$, $\boldsymbol{\xi}$, $\mathbf{w}$, y, $\tilde{u}$, $\tilde{v}$ or σ, we replace $A_q^{[1]}$ by $A_p^{[2]}$. With these modifications, it is easy to check the proof in Sections 3.5–10 applies to Case 2 as well. The main point is we can still gain the factor $1/(N\eta)^{n+1}$ from integral of $g(\cdot)$ defined in (3.20) (with $y_q^{[1]}$ and $\mathbf{w}_q^{[1]}$ replaced by $y_p^{[2]}$ and $\mathbf{w}_p^{[2]}$). Heuristically, we can go back to (4.6), and replace $\sigma_q^{[1]}$ by $\sigma_p^{[2]}$ therein. It is then quite clear the same estimate holds. Consequently, Lemmas 5.1, 5.6, 5.8 and 5.9 still hold under the replacement of the variables described above. Hence, (1.19) holds in Case 2.

• (Case 3)

Analogously, in this case, we can also mimic the discussions for Case 1 with slight modifications. We also start from (3.12) and (3.13). For $A = \breve{\mathcal{S}}$, $\breve{X}$, $\breve{Y}$, $\breve{\Omega}$, $\breve{\Xi}$, $\boldsymbol{\omega}$ and $\boldsymbol{\xi}$, we replace $A_q^{\langle 1\rangle}$ by A_q, and replace $A_q^{[1]}$ by 0. In addition, in the first line of (3.13), we replace $\bar{\phi}_{1,q,1}\phi_{1,p,1}\bar{\phi}_{2,p,1}\phi_{2,q,1}$ by $\bar{\phi}_{1,p,1}\phi_{1,p,1}\bar{\phi}_{2,p,1}\phi_{2,p,1}$. Consequently, after using superbosonization formula, we will get the factor $(y_p^{[1]}|(\mathbf{w}_p^{[1]}(\mathbf{w}_p^{[1]})^*)_{12}|)^{2n}$ instead of $(y_p^{[1]}y_q^{[1]}(\mathbf{w}_q^{[1]}(\mathbf{w}_q^{[1]})^*)_{12}(\mathbf{w}_p^{[1]}(\mathbf{w}_p^{[1]})^*)_{21})^n$ in (3.16). Then, for the superdeterminant terms

$$\prod_{k=p,q}\frac{\det(X_k - \Omega_k(Y_k)^{-1}\Xi_k)}{\det Y_k}, \quad \prod_{k=p,q}\frac{y_k^{[1]}\Big(y_k^{[1]} - \boldsymbol{\xi}_k^{[1]}(X_k^{[1]})^{-1}\boldsymbol{\omega}_k^{[1]}\Big)^2}{\det^2(X_k^{[1]})}.$$

we shall only keep the factors with $k = p$ and delete those with $k = q$. Moreover, we shall also replace $A_q^{[1]}$ by 0 for $A = X$, Y, Ω, Ξ, $\boldsymbol{\omega}$, $\boldsymbol{\xi}$, $\mathbf{w}$, y, $\tilde{u}$, $\tilde{v}$ or σ in (3.16). In addition, $\mathrm{dA}^{[1]}$ shall be redefined as the differential of $A_p^{[1]}$-variables only, for $A = X$, $\mathbf{y}$, $\mathbf{w}$, $\boldsymbol{w}$ and $\boldsymbol{\xi}$. One can check step by step that such a modification does not require any essential change of our discussions for Case 1. Especially,

note that our modification has nothing to do with the saddle point analysis on the Gaussian measure $\exp\{-M(K(\hat{X},V)+L(\hat{B},T))\}$. Moreover, the term $\mathcal{P}(\cdot)$ in (3.28) can be redefined by deleting the factor with $k=q$ in the last term therein. Such a modification does not change our analysis of $\mathcal{P}(\cdot)$. In addition, the irrelevant term $\mathcal{Q}(\cdot)$ can also be defined accordingly. Specifically, we shall delete the factor with $k=q$ in the last term of (3.29) and replace $A_q^{[1]}$ by 0 for $A=\Omega$, Ξ, $\boldsymbol{\omega}$, $\boldsymbol{\xi}$, $\mathbf{w}$, y. It is routine to check that Lemma 6.3 still holds under such a modification. Analogously, we can redefine the functions $\mathcal{F}(\cdot)$, $f(\cdot)$ and $g(\cdot)$ in (3.18)-(3.20). Now, the main difference between Case 3 and Case 1 or 2 is that the factor $(y_p^{[1]}|(\mathbf{w}_p^{[1]}(\mathbf{w}_p^{[1]})^*)_{12}|)^{2n}$ does not produce oscillation in the integral of $g(\cdot)$ any more. Heuristically, the counterpart of (4.6) in Case 3 reads

$$\begin{aligned} &e^{(a_+-a_-)N\eta}\int d\mathbf{y}^{[1]}\mathrm{d}\mathbf{w}^{[1]}d\nu(Q_1)\cdot g(\hat{B},T,Q_1,\mathbf{y}^{[1]},\mathbf{w}^{[1]})\\ &\sim\int_0^\infty 2tdt\int_{\mathbb{L}}d\sigma_p^{[1]}\cdot e^{-cN\eta t^2+c_1e^{-\mathrm{i}\sigma_p^{[1]}}t}\sim\frac{1}{N\eta}. \end{aligned}$$

Hence, (1.19) holds for Case 3.

Therefore, we completed the proof of Theorem 1.14.

12. Further comments

In this section, we make some comments on possible further improvements on our results.

• (Comment on how to remove the prefactor N^{C_0} in (1.19))

As mentioned in Remark 1.15, we have used N^{C_0} to replace $M\Theta^2W^{C_0}/(N\eta)^\ell$. However, the latter is also artificial. It can be improved to some n-dependent constant C_n via a more delicate analysis on $\mathsf{A}(\cdot)$, i.e. the integral of $\mathcal{P}(\cdot)\mathcal{Q}(\cdot)\mathcal{F}(\cdot)$. Such an improvement stems from the cancellation in the Gaussian integral. At first, a finer analysis will show that the factor $\mathcal{Q}(\cdot)$ can really be ignored, in the sense that it does not play any role in the estimate of the order of $\mathbb{E}|G_{ij}(z)|^{2n}$. Hence, for simplicity, we just focus on the product $\mathsf{P}(\cdot)\mathsf{F}(\cdot)$ instead of $\mathsf{A}(\cdot)$. Then, we go back to Lemma 9.2 and Lemma 9.7. Recall the decomposition (9.37). A more careful analysis on $\mathsf{F}(\cdot)$ leads us to the following expansion, up to the subleading order terms of the factors $\mathring{\mathbb{G}}(\cdot)$ and $\mathring{\mathbb{F}}(\cdot)$,

$$\begin{aligned} \mathsf{F}(\cdot)=\mathring{\mathbb{G}}(\cdot)\mathring{\mathbb{F}}(\cdot)\sim&\frac{1}{(N\eta)^{n+2}}\Big(1+\frac{M\eta}{\sqrt{M}}\sum_{j=1}^W\mathsf{l}_j(\mathring{x}_{j,1},\mathring{x}_{j,2},\mathring{v}_j)+\cdots\Big)\\ &\times\Big(1+\frac{M\eta}{\sqrt{M}}\sum_{j=1}^W\mathsf{l}'_j(\mathring{b}_{j,1},\mathring{b}_{j,2},\mathring{t}_j)+\cdots\Big), \end{aligned}\tag{12.1}$$

where $\mathsf{l}_j(\cdot)$'s and $\mathsf{l}'_j(\cdot)$'s are some linear combinations of the arguments. Analogously, we shall write down the leading order term of $\mathsf{P}(\cdot)$ in terms of $\mathring{\mathbf{x}}$, $\mathring{\mathbf{b}}$, $\mathring{\mathbf{t}}$ and $\mathring{\mathbf{v}}$ explicitly, instead of bounding it crudely by using (9.24). Then it can be seen that the leading order term of $\mathsf{P}(\cdot)$ is a linear combination of $\mathring{x}_{j,1}\mathring{x}_{k,2}$, $\mathring{b}_{j,1}\mathring{b}_{k,2}$, $\mathring{x}_{j,\alpha}\mathring{b}_{k,\beta}$, $\upsilon_{j,\alpha}\tau_{k,\beta}$ for $j,k=1,\dots,W$ and $\alpha,\beta=1,2$, in which all the coefficients are of order $1/M$. Observe that the Gaussian integral in (8.47) will kill the linear terms. Consequently, in the expansion (12.1), the first term that survives after the Gaussian integral is actually

$$\frac{1}{(N\eta)^{n+2}}\cdot\frac{M\eta}{\sqrt{M}}\sum_{j=1}^W\mathsf{l}_j(\mathring{x}_{j,1},\mathring{x}_{j,2},\mathring{v}_j)\cdot\frac{M\eta}{\sqrt{M}}\sum_{j=1}^W\mathsf{l}'_j(\mathring{b}_{j,1},\mathring{b}_{j,2},\mathring{t}_j).\tag{12.2}$$

Replacing $\mathsf{A}(\cdot)$ by the product of the leading order term of $\mathsf{P}(\cdot)$ and (12.2) in the integral (8.47) and taking the Gaussian integral over $\mathbf{c}$, $\mathbf{d}$, $\boldsymbol{\tau}$ and $\boldsymbol{\upsilon}$-variables yield the true order $1/(N\eta)^n$, without additional N-dependent prefactors.

• (Comment on the restriction $|E| \le \sqrt{2} - \kappa$)

This restriction is used in several places, we mention the two most important ones.

The most critical issue is the term $\ell_S(\hat{B}, T)$ defined in (5.6). A direct consequence of $|E| \le \sqrt{2}-\kappa$ is that $\mathsf{Re}(b_{j,1} + b_{j,2})(b_{k,1} + b_{k,2}) \ge 0$ for $\mathbf{b}_1 \in \Gamma^W$ and $\mathbf{b}_2 \in \bar{\Gamma}^W$, thus $\mathsf{Re}\ell_S(\hat{B}, T) \ge 0$. However, once $|E| > \sqrt{2}$, $\mathsf{Re}\ell_S(\hat{B}, T)$ can be negative for $\mathbf{b}_1 \in \Gamma^W$ and $\mathbf{b}_2 \in \bar{\Gamma}^W$. Consequently, the measure $\exp\{-M\ell_S(\hat{B}, T)\}$ is not well defined, considering the domain of the $\mathbf{t}$-variables is not compact. Actually, such a problem is unavoidable if we independently deform the contours of $b_{j,1}$'s (resp. $b_{j,2}$'s) from $\mathbb{R}_+$ to any contour passing through the saddle point a_+ (resp. a_-), starting from the 0, since one can always choose $b_{j,1}$ and $b_{k,1}$ to be a_+ while $b_{j,2}$ and $b_{k,2}$ to be very close to 0, such that $\mathsf{Re}(b_{j,1} + b_{j,2})(b_{k,1} + b_{k,2}) < 0$ in case $|E| > \sqrt{2}$. A possible way to solve this problem is to change the variables $(b_{j,1}, b_{j,2})$ to $(\mathsf{r}_{j,1}, \mathsf{r}_{j,2})$, defined by $\mathsf{r}_{j,1} = (b_{j,1} + b_{j,2})/2$ and $\mathsf{r}_{j,2} = (b_{j,1} - b_{j,2})/2$, and perform the saddle point analysis with respect to the latter. It is not difficult to calculate that the saddle point of $L(\hat{B}, T)$ restricted on these r-variables is $\mathsf{r}_{j,1} = \sqrt{4 - E^2}$ and $\mathsf{r}_{j,2} = \mathbf{i}E$. To guarantee the positivity of $\ell_S(\hat{B}, T)$, it suffices to choose a contour for $\mathsf{r}_{j,1}$ passing through $\sqrt{4 - E^2}$, staying in the sector $\tilde{\mathbb{K}} := \{\omega \in \mathbb{C} : |\arg \omega| < \pi/4\}$. However, the explicit way to choose the contours for $\mathsf{r}_{j,1}$ and $\mathsf{r}_{j,2}$ and the subsequent analysis in terms of r-variables instead of $\mathbf{b}$-variables would be more involved and we leave it to future work.

Another point where $|E| < \sqrt{2} - \kappa$ is used is the proof of Lemma 5.9 in Section 10. In (10.3), we used $\mathsf{Re}\mathbb{A}_+^{\upsilon} \ge I$, which is a consequence of $\mathsf{Re}a_+^2 = (2 - E^2)/2 \ge 0$ and (5.23), see the definition of $\mathbb{A}_+^{\upsilon}$ in (8.35). If $|E| > \sqrt{2}$, the bound in (10.3) should be weakened to $\mathsf{Re}(\mathring{\mathbf{x}}'\mathbb{A}_+^{\upsilon}\mathring{\mathbf{x}}) \ge c(E)||\mathring{\mathbf{x}}||_2^2$ for some E-dependent constant $c(E) < 1$, which is uniform on $V_j \in \mathring{U}(2)$ for all $j = 2, \ldots, W$. This modification, however, is not critical. An easy way to remedy this situation is to impose an additional condition $\det S^{(1)} \le (c(E) - c)^W$ for arbitrarily small constant c, to guarantee the estimate (10.7).

References

[1] Aizenman, M., and Molchanov, S.: Localization at large disorder and at extreme energies: an elementary derivation, *Commun. Math. Phys.* **157**, 245–278 (1993)

[2] Anderson, P.: Absences of diffusion in certain random lattices, *Phys. Rev.* **109**, 1492–1505 (1958)

[3] Bunder J.E., Efetov K.B., Kravtsov V.E., Yevtushenko O.M. and Zirnbauer M.R.: Superbosonization formula and its application to random matrix theory. *J. Stat. Phys.* **129** (5–6), 809–832 (2007)

[4] Disertori, M., Pinson, H., Spencer, T.: Density of states for random band matrices. *Commun. Math. Phys.* **232**(1), 83–124 (2002)

[5] Disertori, M., Spencer, T.: Anderson localization for a supersymmetric sigma model. *Commun. Math. Phys.* **300**(3), 659–671 (2010)

[6] Disertori, M., Spencer, T., Zirnbauer, M.R.: Quasi-diffusion in a 3D supersymmetric hyperbolic sigma model. *Commun. Math. Phys.* **300**(2), 435–486 (2010)

[7] Efetov, K.: Supersymmetry in disorder and chaos. Cambridge University Press, 1997.

[8] Ellis, R. B.: Discrete Green's functions for products of regular graphs. arXiv preprint math/0309080, (2003)

[9] Erdős, L., Knowles, A., Yau, H.-T.: Averaging fluctuations in resolvents of random band matrices. *Ann. Henri Poincaré* **14**(8), 1837–1926 (2013)

[10] Erdős, L., Knowles, A., Yau, H.-T., and Yin, J.: Spectral statistics of Erdős-Rényi graphs I: local semicircle law. *Ann. Prob.* **41**(3B), 2279–2375 (2013)

[11] Erdős, L., Knowles, A., Yau, H.-T., Yin, J.: Delocalization and Diffusion Profile for Random Band Matrices. *Commun. Math. Phys.* **323**, 367–416 (2013)

[12] Erdős, L., Knowles, A., Yau, H.-T., Yin, J.: The local semicircle law for a general class of random matrices. *Electron. J. Probab,* **18**(59), 1–58 (2013)

[13] Erdős, L., Schlein, B., Yau, H.-T.: Local semicircle law and complete delocalization for Wigner random matrices. *Commun. Math. Phys.* **287**, 641–655 (2009)

[14] Erdős, L., Yau, H.-T., Yin, J.: Bulk universality for generalized Wigner matrices. *Probab. Theory and Related Fields* **154**(1-2), 341–407 (2012)

[15] Fröhlich, J., Spencer, T.: Absence of diffusion in the Anderson tight binding model for large disorder or low energy, *Commun. Math. Phys.* **88**, 151–184 (1983)

[16] Fyodorov, Y.V. and Mirlin, A.D.: Scaling properties of localization in random band matrices: a σ-model approach. *Phys. Rev. Lett.* **67**, 2405–2409 (1991)

[17] Littelmann, P., Sommers, H.J., Zirnbauer, M.R.: Superbosonization of invariant random matrix ensembles. *Commun. Math. Phys.* **283**(2), 343–395 (2008)

[18] Shcherbina, T.: On the second mixed moment of the characteristic polynomials of 1D band matrices. *Commun. Math. Phys.* **328**(1), 45–82 (2014)

[19] Shcherbina, T.: Universality of the local regime for the block band matrices with a finite number of blocks. *J. Stat. Phys.* **155**(3), 466–499 (2014)

[20] Shcherbina, T.: Universality of the second mixed moment of the characteristic polynomials of the 1D band matrices: real symmetric case. arXiv:1410.3084, 2014.

[21] Schäfer, L., Wegner, F.: Disordered system with n orbitals per site: Lagrange formulation, hyperbolic symmetry, and Goldstone modes. *Z. Phys. B* **38**, 113–126 (1980)

[22] Schenker, J.: Eigenvector localization for random band matrices with power law band width. *Commun. Math. Phys.* **290**, 1065–1097 (2009)

[23] Sodin, S.: An estimate for the average spectral measure of random band matrices, *J. Stat. Phys.* **144** (2011), no. 1, 46-59.

[24] Sodin, S.: The spectral edge of some random band matrices. *Ann. of Math.* **172**(3), 2223-2251 (2010)

[25] Spencer, T.: SUSY statistical mechanics and random band matrices. Quantum many body systems. Springer Berlin Heidelberg, 125-177 (2012)

[26] Spencer, T, Zirnbauer, M.R.: Spontaneous symmetry breaking of a hyperbolic sigma model in three dimensions. *Commun. Math. Phys.* **252**(1-3), 167–187 (2004)

[27] Tao, T., Vu, V.: Random matrices: universality of local eigenvalue statistics. *Acta Math.* **206**(1), 127–204 (2011)

[28] Wegner, F.J.: Disordered system with n orbitals per site: $n \to \infty$ limit, *Phys. Rev. B* **19**, 783-792 (1979)

[29] Wigner, E.: Characteristic vectors of bordered matrices with infinite dimensions. *Ann. of Math.* **62**, 548–564 (1955)

IST Austria, Am Campus 1, 3400 Klosterneuberg, Austria
E-mail address: maomie2007@gmail.com

IST Austria, Am Campus 1, 3400 Klosterneuberg, Austria
E-mail address: laszlo.erdos@ist.ac.at